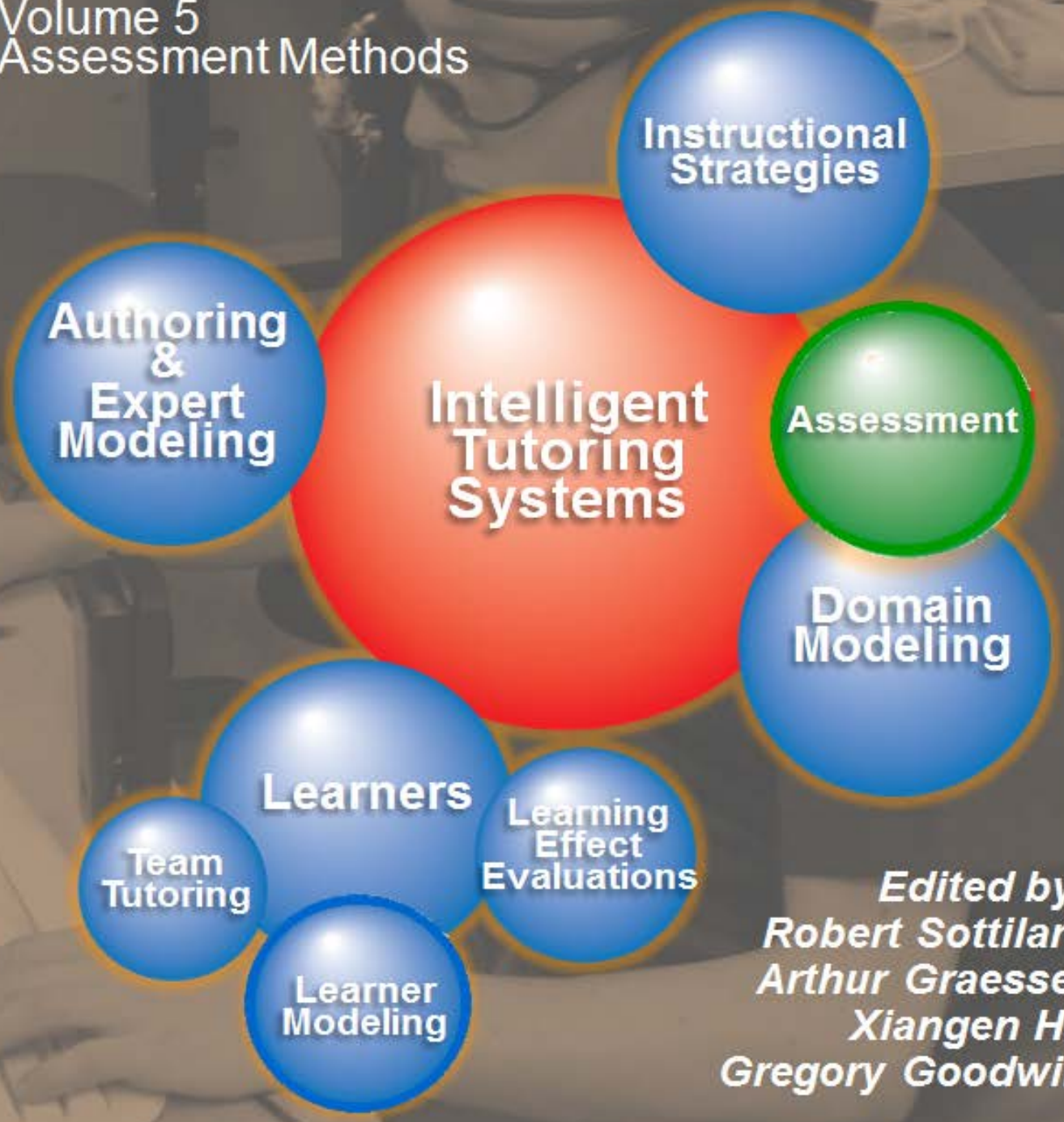
Design Recommendations for Intelligent Tutoring Systems

Volume 5
Assessment Methods



Edited by:
Robert Sottilare
Arthur Graesser
Xiangen Hu
Gregory Goodwin

A Book in the Adaptive Tutoring Series

Design Recommendations for Intelligent Tutoring Systems

Volume 5
Assessment Methods

Edited by:
Robert A. Sottilare
Arthur C. Graesser
Xiangen Hu
Gregory A. Goodwin

A Book in the Adaptive Tutoring Series

Printed in the United States of America
First Printing, August 2017

US Army Research Laboratory
Human Research & Engineering Directorate
Orlando, Florida

International Standard Book Number: 978-0-9977257-2-8

We wish to acknowledge the editing and formatting contributions of Carol Johnson, ARL.

Special thanks to Jody Cockroft, University of Memphis, for her efforts in coordinating the workshop that led to this volume.

Dedicated to current and future scientists and developers of adaptive learning technologies

CONTENTS

INTRODUCTION TO ASSESSMENT METHODS & GIFT **1**

Section I - Competency Assessment 15

CHAPTER 1 – Understanding Competency Assessment Methodologies Applied to Adaptive Instruction in Intelligent Tutoring Systems **17**

Robert Sottilare, Ph.D.

CHAPTER 2 – Competence-based Knowledge Structures and Current Challenges for E-Assessment **21**

Dietrich Albert, Alexander Nussbaumer, Bor-Chen Kuo, Peter W. Foltz, and Xiangen Hu

CHAPTER 3 – Exploring Assessment Mechanisms in the Total Learning Architecture (TLA) **29**

Gregory Goodwin, J. T. Folsom-Kovarik, Andy Johnson, Sae Schatz, and Robert Sottilare

CHAPTER 4 – Enabling Intelligent Tutoring System Tracking with the Experience Application Programming Interface (xAPI) **41**

Andy Johnson, Benjamin D Nye, Diego Zapata-Rivera, and Xiangen Hu

CHAPTER 5 – Vision Statement: Navy Career Management and Training of the Future **47**

Brent Olde

CHAPTER 6 – Coordinating Evidence Across Learning Modules Using Digital Badges **53**

Ross Higashi, Christian Schunn, Vu Nguyen, and Scott J. Ososky

CHAPTER 7 – Leveraging Domain Models for Personalizing Problem Solving and Learning **69**

Louise Yarnall, Eric Snow, Erica Snow, and Irvin R. Katz

CHAPTER 8 – Assessing Individual Learner Performance in MOOCs **85**

Ryan S. Baker, Piotr Mitros, Benjamin Goldberg, and Robert A. Sottilare

Section II - Evidence-Centered Design and Data Mining 97

CHAPTER 9 – Evidence Centered Design and Data-Driven Assessment **99**

Arthur C. Graesser

CHAPTER 10 – Evidence-Centered Assessment Design and Probability-Based Inference to Support the Generalized Intelligent Framework for Tutoring (GIFT) **101**

Robert J. Mislevy and Duanli Yan

CHAPTER 11 – Reusing Evidence in Assessment and Intelligent Tutors **125**

Diego Zapata-Rivera, Keith Brawner, G. Tanner. Jackson, and Irvin R. Katz

CHAPTER 12 – Methods for Assessing Inquiry: Machine-learned and Theoretical **137**

Michelle LaMar, Ryan S. Baker, and Samuel Greiff

CHAPTER 13 – Automated Assessment of Learner-Generated Natural Language Responses **155**

Vasile Rus, Andrew M. Olney, Peter W. Foltz, and Xiangen Hu

CHAPTER 14 – Using Process Data for Assessment in Intelligent Tutoring Systems: A Cognitive Psychologist, Psychometrician, and Computer Scientist Perspective **171**

Samuel Greiff, Dragan Gašević, and Alina A. von Davier

Section III - General Assessment Methods 181

CHAPTER 15 – Principles of Assessment in the Generalized Intelligent Framework for Tutoring (GIFT) **183**

Gregory A. Goodwin

CHAPTER 16 – Why Assess? The Role of Assessment in Learning Science and Society 189

Benjamin D. Nye, Piotr Mitros, Christian Schunn, Peter W. Foltz, Dragan Gašević, and Irvin R. Katz

CHAPTER 17 – Assessment of Forgetting 203

Philip I. Pavlik Jr., Jaclyn K. Maass, and Jong W. Kim

CHAPTER 18 – Validity Issues and Concerns for Technology-based Performance Assessments 209

Irvin R. Katz, Michelle M. LaMar, Randall Spain, Juan Diego Zapata-Rivera, Jo-Anne Baird, and Samuel Greiff

CHAPTER 19 – Toward Systematic Assessment of Human Performance Interventions in the US Army: An Assessment Process Framework 225

Kara L. Orvis, Jared T. Freeman, Jeffrey M. Beaubien, Clayton W. Burford, Joan H. Johnston, Lauren Reinerman-Jones, and Grace Teo

CHAPTER 20 – Assessment in Intelligent Tutoring Systems in Traditional, Mixed Mode, and Online Courses 235

Anne M. Sinatra, Scott Ososky, and Robert Sottilare

CHAPTER 21 – Lessons Learned from Large-Scale E-Assessments: Future Directions for the Generalized Intelligent Framework for Tutoring (GIFT) 249

Jo-Anne Baird, Anne M. Sinatra, and Gregory Goodwin

Section IV - Assessment Methods for Particular Domains and Problems 257

CHAPTER 22 – Selected Assessment Techniques for the Generalized Intelligent Framework for Tutoring (GIFT) 259

Xiangen Hu

CHAPTER 23 – Assessing Teacher Questions in Classrooms 261

Andrew M. Olney, Sean Kelly, Borhan Samei, Patrick Donnelly, and Sidney K. D'Mello

CHAPTER 24 – Assessment of Collaborative Problem Solving 275

Arthur C. Graesser, Zhiqiang Cai, Xiangen Hu, Peter W. Foltz, Samuel Greiff, Bor-Chen Kuo, Chen- Huei Liao, and David Williamson Shaffer

CHAPTER 25 – Challenges for Assessing and Tutoring Collective Skills 287

Jeanine Ayers, Martin L. Bink, and Frederick J. Diedrich

CHAPTER 26 – Cognitive Assessment as Service in the Generalized Intelligent Framework for Tutoring (GIFT) 295

Xiangen Hu, Sheng Xu, Robert Sottilare, and Dietrich Albert

CHAPTER 27 – Assessment in AutoTutor 309

Zhiqiang Cai, Arthur C. Graesser, Xiangen Hu, and Bor-chen Kuo

CHAPTER 28 – Assessment of Individual Learner Performance in Psychomotor Domains 319

Jong W. Kim, Robert A. Sottilare, Gregory Goodwin, and Xiangen Hu

CHAPTER 29 – Motivating Individual Difference in an Intelligent Tutoring System 331

Lauren Reinerman-Jones, Elizabeth Lameier, Elizabeth Biddle, and Michael Boyce

Biographies 341

Index 353

INTRODUCTION TO ASSESSMENT METHODS & GIFT

Robert A. Sottilare[1], Arthur C. Graesser[2], Xiangen Hu[2], and Gregory A. Goodwin[1] , Eds.

U.S. Army Research Laboratory - Human Research and Engineering Directorate[1]
University of Memphis Institute for Intelligent Systems[2]

This book is the fifth in a planned series of books that examine key topics (e.g., learner modeling, instructional strategies, authoring, domain modeling, assessment, impact on learning, team tutoring, machine learning, and potential standards) in intelligent tutoring system (ITS) design through the lens of the Generalized Intelligent Framework for Tutoring (GIFT) (Sottilare, Brawner, Goldberg & Holden, 2012; Sottilare, Brawner, Sinatra, & Johnston, 2017). GIFT is a modular, service-oriented architecture created to reduce the cost and skill required to author ITSs, manage instruction within ITSs, and evaluate the effect of ITS technologies on learning, performance, retention, transfer of skills, and other instructional outcomes.

Along with this volume, the first four books in this series, Learner Modeling (ISBN 978-0-9893923-0-3), Instructional Management (ISBN 978-0-9893923-2-7), Authoring Tools (ISBN 978-0-9893923-6-5) and Domain Modeling (978-0-9893923-9-6) are freely available at www.GIFTtutoring.org and on Google Play.

This introduction begins with a description of tutoring functions, provides a glimpse of assessment best practices, and examines the motivation for standards in the design, authoring, instruction, and evaluation of ITS tools and methods. We introduce GIFT design principles and discuss how readers might use this book as a design tool. We begin by examining the major components of ITSs.

Components and Functions of Intelligent Tutoring Systems

It is generally accepted that an ITS has four major components (Elson-Cook, 1993; Nkambou, Mizoguchi & Bourdeau, 2010; Graesser, Conley & Olney, 2012; Psotka & Mutter, 2008; Sleeman & Brown, 1982; VanLehn, 2006; Woolf, 2009): the domain model, the student model, the tutoring model, and the user-interface model. GIFT similarly adopts this four-part distinction, but with slightly different corresponding labels (domain module, learner module, pedagogical module, and tutor-user interface) and the addition of the sensor module, which can be viewed as an expansion of the user interface.

(1) The **domain model** contains the set of skills, knowledge, and strategies/tactics of the topic being tutored. It normally contains the ideal expert knowledge and also the bugs, mal-rules, and misconceptions that students periodically exhibit.

(2) The **learner model** consists of the cognitive, affective, motivational, and other psychological states that evolve during the course of learning. Since learner performance is primarily tracked in the domain model, the learner model is often viewed as an overlay (subset) of the domain model, which changes over the course of tutoring. For example, "knowledge tracing" tracks the learner's progress from problem to problem and builds a profile of strengths and weaknesses relative to the domain model (Anderson, Corbett, Koedinger & Pelletier, 1995). An ITS may also consider psychological states outside of the domain model that need to be considered as parameters to guide tutoring.

(3) The **tutor model** (also known as the pedagogical model or the instructional model) takes the domain and learner models as input and selects tutoring strategies, steps, and actions on what the tutor should do next in the exchange. In mixed-initiative systems, the learners may also take actions, ask questions, or request help (Aleven, McClaren, Roll & Koedinger, 2006; Rus & Graesser, 2009), but the ITS always needs to be ready to decide "what to do next" at any point and this is determined by a tutoring model that captures the researchers' pedagogical theories.

(4) The **user interface** interprets the learner's contributions through various input media (speech, typing, clicking) and produces output in different media (text, diagrams, animations, agents). In addition to the conventional human-computer interface features, some recent systems have incorporated natural language interaction (Graesser et al., 2012; Johnson & Valente, 2008), speech recognition (D'Mello, Graesser & King, 2010; Litman, 2013), and the sensing of learner emotions (Baker,

D'Mello, Rodrigo & Graesser, 2010; D'Mello & Graesser, 2010; Goldberg, Sottilare, Brawner, Holden, 2011).

The designers of a tutor model must make decisions on each of the various major components in order to create an enhanced learning experience through well-grounded pedagogical strategies (optimal plans for action by the tutor) that are selected based on learner states and traits and that are delivered to the learner as instructional tactics (optimal actions by the tutor). Next, tactics are chosen based on the previously selected strategies and instructional context (the conditions of the training at the time of the instructional decision). This is part of the learning effect model (Sottilare, 2012; Fletcher & Sottilare, 2013; Sottilare, 2013; Sottilare, Ragusa, Hoffman & Goldberg, 2013), which has been updated and described below in more detail in the section titled "Motivations for Intelligent Tutoring System Standards" in this introductory chapter.

Principles of Learning and Instructional Techniques, Strategies, and Tactics

Instructional techniques, strategies, and tactics play a central role in the design of GIFT. Instructional techniques represent instructional best practices and principles from the literature, many of which have yet to be implemented within GIFT at the writing of this volume. Examples of instructional techniques include, but are not limited to, error-sensitive feedback, mastery learning, adaptive spacing and repetition, and fading worked examples. Others are represented in the next section of this introduction. It is anticipated that techniques within GIFT will be implemented as software-based agents where the agent will monitor learner progress and instructional context to determine if best practices (agent policies) have been adhered to or violated. Over time, the agent will learn to enforce agent policies in a manner that optimizes learning and performance.

Some of the best instructional practices (techniques) have yet to be implemented in GIFT, but many instructional strategies and tactics have been implemented. Instructional strategies (plans for action by the tutor) are selected based on changes to the learner's state (cognitive, affective, physical). If a sufficient change in any learner's state occurs, this triggers GIFT to select a generic strategy (e.g., provide feedback). The instructional context along with the instructional strategy then triggers the specific selection of an instructional tactic (an action to be taken by the tutor). If the strategy is to "provide feedback," then the tactic might be to "provide feedback on the error committed during the presentation of instructional concept 'B' in the chat window during the next turn." Tactics detail what is to be done, why, when, and how.

An adaptive, intelligent learning environment needs to select the right instructional strategies at the right time, based on its model of the learner in specific conditions and the learning process in general. Such selections should be taken to maximize deep learning and motivation while minimizing training time and costs.

Motivations for Intelligent Tutoring System Standards

An emphasis on self-regulated learning has highlighted a requirement for point-of-need training in environments where human tutors are either unavailable or impractical. ITSs have been shown to be as effective as expert human tutors (VanLehn, 2011) in one-to-one tutoring in well-defined domains (e.g., mathematics or physics) and significantly better than traditional classroom training environments. ITSs have demonstrated significant promise, but 50 years of research have been unsuccessful in making ITSs ubiquitous in military training or the tool of choice in our educational system. This begs the question: "Why?"

Part of the answer lies in the fact that the availability and use of ITSs have been constrained by their high development costs, their limited reuse, a lack of standards, and their inadequate adaptability to the needs of learners. Educational and training technologies like ITSs are primarily researched and developed in a few key environments: industry, academia, and government including military domains. Each of these environments has its own challenges and design constraints. The application of ITSs to military domains is further hampered by the complex and often ill-defined environments in which the US military operates today. ITSs are often built as domain-specific, unique, one-of-a-kind, largely domain-dependent solutions focused on a single pedagogical strategy (e.g., model tracing or constraint-based approaches) when complex learning domains may require novel or hybrid approaches. Therefore, a modular ITS framework and standards are needed to enhance reuse, support authoring, optimize instructional strategies, and lower the cost and skillset needed for users to adopt ITS solutions for training and education. It was out of this need that the idea for GIFT arose.

GIFT has three primary functions: authoring, instructional management, and evaluation. First, it is a framework for authoring new ITS components, methods, strategies, and whole tutoring systems. Second, GIFT is an instructional manager that integrates selected instructional theory, principles, and strategies for use in ITSs. Finally, GIFT is an experimental testbed used to evaluate the effectiveness and impact of ITS components, tools, and methods. GIFT is based on a learner-centric approach with the goal of improving linkages in the updated adaptive tutoring learning effect model (Figure 1; Sottilare, Burke, Salas, Sinatra, Johnston, & Gilbert, 2017).

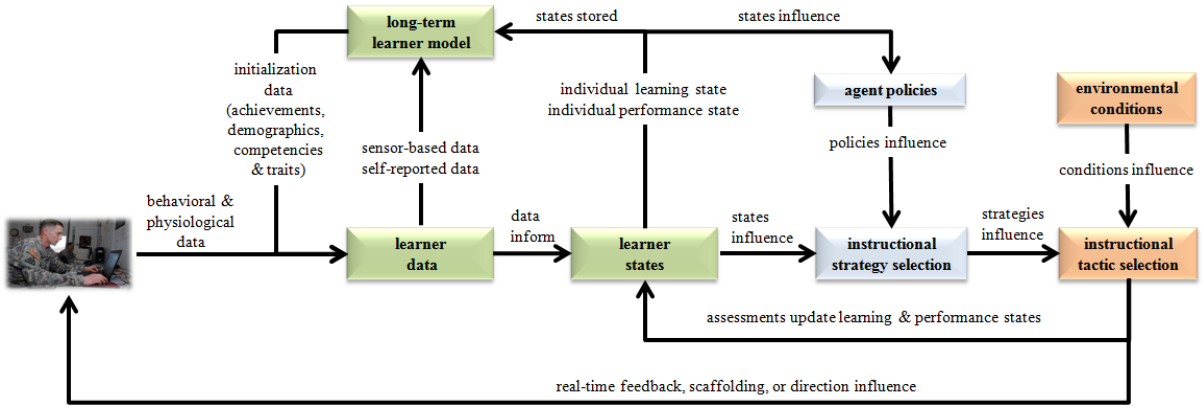


Figure 1. Updated adaptive tutoring learning effect model (Sottilare et al, 2017)

A deeper understanding of the learner's behaviors, traits, and preferences (learner data) collected through performance, physiological and behavioral sensors, and surveys will allow for more accurate evaluation of the learner's states (e.g., engagement level, confusion, frustration). This will result in a better and more persistent model of the learner. To enhance the adaptability of the ITS, methods are needed to accurately classify learner states (e.g., cognitive, affective, psychomotor, social) and select optimal instructional strategies given the learner's existing states. A more comprehensive learner model will allow the ITS to adapt more appropriately to address the learner's needs by changing the instructional strategy (e.g., content, flow, or feedback). An instructional strategy better aligned to the learner's needs is more likely to positively influence their learning gains. It is with the goal of optimized learning gains in mind that the design principles for GIFT were formulated.

This version of the learning effect model has been updated to gain understanding of the effect of optimal instructional tactics and instructional context (both part of the domain model) on specific desired outcomes including knowledge and skill acquisition, performance, retention, and transfer of skills from training or tutoring environments to operational contexts (e.g., from practice to application). The feedback loops in

Figure 1 have been added to identify tactics as either a change in instructional context or interaction with the learner. This allows the ITS to adapt to the need of the learner. Consequently, the ITS changes over time by reinforcing learning mechanisms.

GIFT Design Principles

The GIFT methodology for developing a modular, computer-based tutoring framework for training and education considered major design goals, anticipated uses, and applications. The design process also considered enhancing one-to-one (individual) and one-to-many (collective or team) tutoring experiences beyond the state of practice for ITSs today. A significant focus of the GIFT design was on domain-dependent elements in the domain module only. This is a design tradeoff to foster reuse and allows ITS decisions and actions to be made across any/all domains of instruction.

One design principle adopted in GIFT is that each module should be capable of gathering information from other modules according to the design specification. Designing to this principle resulted in standard message sets and message transmission rules (i.e., request-driven, event-driven, or periodic transmissions). For instance, the pedagogical module is capable of receiving information from the learner module to develop courses of action for future instructional content to be displayed, manage flow and challenge level, and select appropriate feedback. Changes to the learner's state (e.g., engagement, motivation, or affect) trigger messages to the pedagogical module, which then recommends general courses of action (e.g., ask a question or prompt the learner for more information) to the domain module, which provides a domain-specific intervention (e.g., what is the next step?).

Another design principle adopted within GIFT is the separation of content from the executable code (Patil & Abraham, 2010). Data and data structures are placed within models and libraries, while software processes are programmed into interoperable modules. Efficiency and effectiveness goals (e.g., accelerated learning and enhanced retention) were considered to address the time available for military training and the renewed emphasis on self-regulated learning. An outgrowth of this emphasis on efficiency and effectiveness led Dr. Sottilare to seek external collaboration and guidance. In 2012, ARL with the University of Memphis developed expert workshops of senior tutoring system scientists from academia and government to influence the GIFT design goals moving forward. Expert workshops have been held each year since 2012 resulting in volumes in the Design Recommendations for Intelligent Tutoring Systems series the following year. The learner modeling expert workshop was completed in September 2012 and Volume 1 followed in July 2013. An expert workshop on instructional management was completed in July 2013 and Volume 2 followed in June 2014. The authoring tools expert workshop was completed in June of 2014 and Volume 3 was published in June 2015.The domain modeling expert workshop was held in June 2015 and Volume 4 was published in July 2016, the assessment expert workshop was held in May 2016, and the team tutoring expert workshop was held in May 2017. Future expert workshops are planned for machine learning techniques, potential standards, and learning effect evaluation methods.

Design Goals and Anticipated Uses

GIFT may be used for a number of purposes, with the primary ones enumerated below:

1. An architectural framework with modular, interchangeable elements and defined relationships to support stand-alone tutoring or guided training if integrated with a training system

2. A set of specifications to guide ITS development

3. A set of exemplars or use cases for GIFT to support authoring, reuse, and ease-of-use

4. A technical platform or testbed for guiding the evaluation, development/refinement of concrete systems

These use cases have been distilled down into the three primary functional areas: authoring, instructional management, and the recently renamed evaluation function. Discussed below are the purposes, associated design goals, and anticipated uses for each of the GIFT functions.

GIFT Authoring Function

The purpose of the GIFT authoring function is to provide technology (tools and methods) to make it affordable and easier to build ITSs and ITS components. Toward this end, a set of authoring interfaces with backend XML configuration tools continues to be developed to allow for data-driven changes to the design and implementation of GIFT-generated ITSs. The design goals for the GIFT authoring function have been adapted from Murray (1999, 2003) and Sottilare and Gilbert (2011). The GIFT authoring design goals are as follow:

- Decrease the effort (time, cost, and/or other resources) for authoring and analyzing ITSs by automating authoring processes, developing authoring tools and methods, and developing standards to promote reuse.
- Decrease the skill threshold by tailoring tools for specific disciplines (e.g., instructional designers, training developers, and trainers) to author, analyze, and employ ITS technologies.
- Provide tools to aid designers/authors/trainers/researchers in organizing their knowledge.
- Support (structure, recommend, or enforce) good design principles in pedagogy through user interfaces and other interactions.
- Enable rapid prototyping of ITSs to allow for rapid design/evaluation cycles of prototype capabilities.
- Employ standards to support rapid integration of external training/tutoring environments (e.g., simulators, serious games, slide presentations, transmedia narratives, and other interactive multimedia).
- Develop/exploit common tools and user interfaces to adapt ITS design through data-driven means.
- Promote reuse through domain-independent modules and data structures.
- Leverage open-source solutions to reduce ITS development and sustainment costs.
- Develop interfaces/gateways to widely-used commercial and academic tools (e.g., games, sensors, toolkits, virtual humans).

As a user-centric architecture, anticipated uses for GIFT authoring tools are driven largely by the anticipated users, which include learners, domain experts, instructional system designers, training and tutoring system developers, trainers and teachers, and researchers. In addition to user models and GUIs, GIFT authoring tools include domain-specific knowledge configuration tools, instructional strategy development tools, and a compiler to generate executable ITSs from GIFT components in a variety of formats (e.g., PC, Android, and iPad).

Within GIFT, domain-specific knowledge configuration tools permit authoring of new knowledge elements or reusing existing (stored) knowledge elements. Domain knowledge elements include learning objectives, media, task descriptions, task conditions, standards and measures of success, common misconceptions, feedback library, and a question library, which are informed by instructional system design principles that, in turn, inform concept maps for lessons and whole courses. The task descriptions, task conditions, standards and measures of success, and common misconceptions may be informed by an expert or ideal learner model derived through a task analysis of the behaviors of a highly skilled user. ARL is investigating techniques to automate this expert model development process to reduce the time and cost of developing ITSs. In addition to feedback and questions, supplementary tools are anticipated to author explanations, summaries, examples, analogies, hints, and prompts in support of GIFT's instructional management function.

GIFT Instructional Management Function

The purpose of the GIFT instructional management function is to integrate pedagogical best practices in GIFT-generated ITSs. The modularity of GIFT will also allow GIFT users to extract pedagogical models for use in tutoring/training systems that are not GIFT-generated. GIFT users may also integrate pedagogical models, instructional strategies, or instructional tactics from other tutoring systems into GIFT. The design goals for the GIFT instructional management function are the following:

- Support ITS instruction for individuals and small teams in local and geographically distributed training environments (e.g., mobile training), and in both well-defined and ill-defined learning domains.

- Provide for comprehensive learner models that incorporate learner states, traits, demographics, and historical data (e.g., performance) to inform ITS decisions to adapt training/tutoring.

- Support low-cost, unobtrusive (passive) methods to sense learner behaviors and physiological measures and use these data along with instructional context to inform models to classify (in near real time) the learner's states (e.g., cognitive and affective).

- Support both macro-adaptive strategies (adaptation based on pre-training learner traits) and micro-adaptive instructional strategies and tactics (adaptation based learner states and state changes during training).

- Support the consideration of individual differences where they have empirically been documented to be significant influencers of learning outcomes (e.g., knowledge or skill acquisition, retention, and performance).

- Support adaptation (e.g., pace, flow, and challenge level) of the instruction based the domain and learning class (e.g., cognitive learning, affective learning, psychomotor learning, social learning).

- Model appropriate instructional strategies and tactics of expert human tutors to develop a comprehensive pedagogical model.

To support the development of optimized instructional strategies and tactics, GIFT is heavily grounded in learning theory, tutoring theory, and motivational theory. Learning theory applied in GIFT includes conditions of learning and theory of instruction (Gagne, 1985), component display theory (Merrill, Reiser, Ranney & Trafton, 1992), cognitive learning (Anderson & Krathwohl, 2001), affective learning (Krathwohl, Bloom & Masia, 1964; Goleman, 1995), psychomotor learning (Simpson, 1972), and social learning (Sottilare, Holden, Brawner, & Goldberg, 2011; Soller, 2001). Aligning with our goal to model expert human

tutors, GIFT considers the intelligent, nurturant, Socratic, progressive, indirect, reflective, and encouraging (INSPIRE) model of tutoring success (Lepper, Drake, & O'Donnell-Johnson, 1997) and the tutoring process defined by Person, Kreuz, Zwaan, and Graesser (1995) in the development of GIFT instructional strategies and tactics.

Human tutoring strategies have been documented by observing tutors with varying levels of expertise. For example, Lepper's INSPIRE model is an acronym that highlights the seven critical characteristics of successful tutors. Graesser and Person's (1994) 5-step tutoring frame is a common pattern of the tutor-learner interchange in which the tutor asks a question, the learner answers the question, the tutor gives short feedback on the answer, then the tutor and learner collaboratively improve the quality of (or embellish) the answer, and finally, the tutor evaluates whether the learner understands the answer. Cade, Copeland, Person, and D'Mello (2008) identified a number of tutoring modes used by expert tutors, which hopefully could be integrated with ITS.

As a learner-centric architecture, anticipated uses for GIFT instructional management capabilities include both automated instruction and blended instruction, where human tutors/teachers/trainers use GIFT to support their curriculum objectives. If its design goals are realized, it is anticipated that GIFT will be widely used beyond military training contexts as GIFT users expand the number and type of learning domains and resulting ITS generated using GIFT.

GIFT Evaluation Function

The GIFT Analysis Function has recently migrated to become the GIFT Evaluation Function with an emphasis on the evaluation of effect on learning, performance, retention and transfer. The purpose of the GIFT evaluation function is to allow ITS researchers to experimentally assess and evaluate ITS technologies (ITS components, tools, and methods). The design goals for the GIFT evaluation function are the following:

- Support the conduct of formative assessments to improve learning.
- Support summative evaluations to gauge the effect of technologies on learning.
- Support assessment of ITS processes to understand how learning is progressing throughout the tutoring process.
- Support evaluation of resulting learning versus stated learning objectives.
- Provide diagnostics to identify areas for improvement within ITS processes.
- Support the ability to comparatively evaluate ITS technologies against traditional tutoring or classroom teaching methods.
- Develop a testbed methodology to support assessments and evaluations (Figure 2).

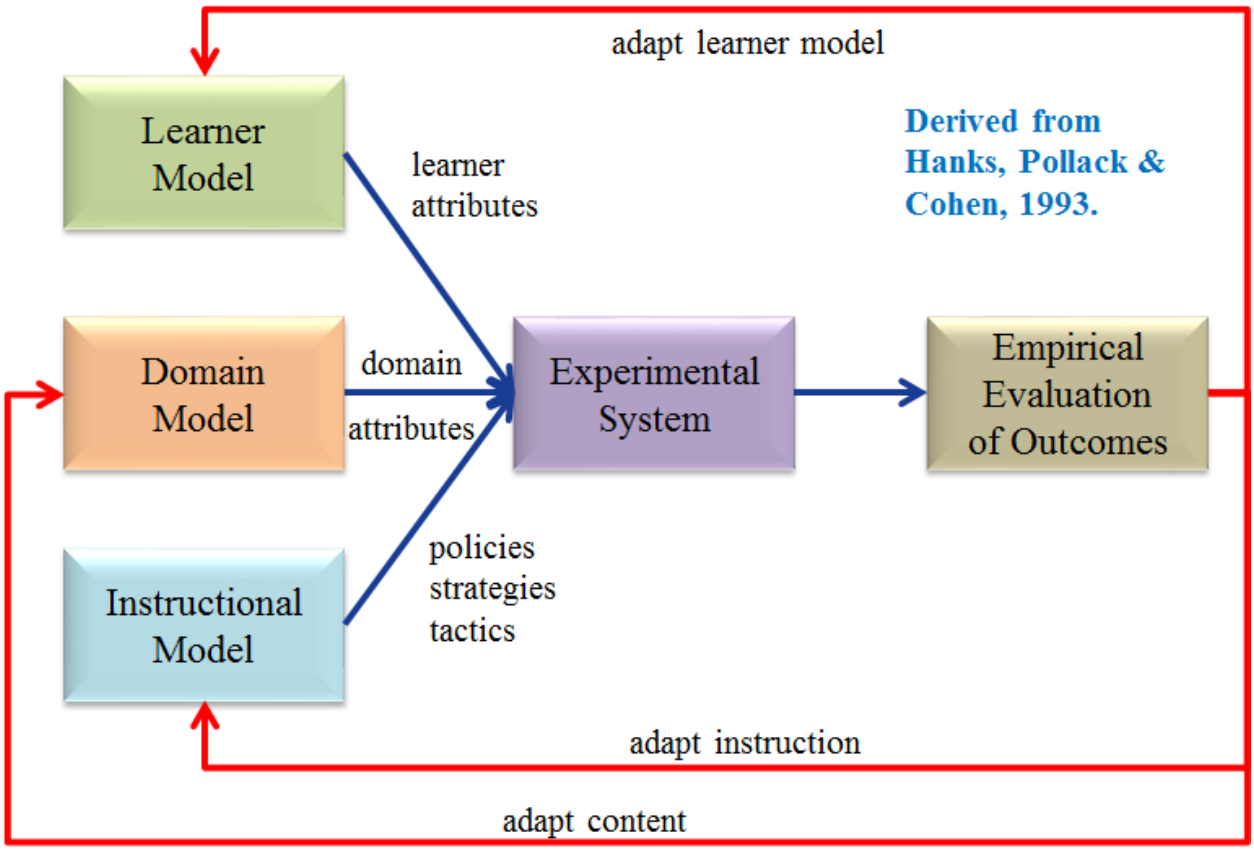


Figure 2. GIFT evaluation testbed methodology

Figure 2 illustrates an analysis testbed methodology being implemented in GIFT. This methodology was derived from Hanks, Pollack, and Cohen (1993). It supports manipulation of the learner model, instructional strategies, and domain-specific knowledge within GIFT, and may be used to evaluate variables in the adaptive tutoring learning effect model (Sottilare, 2012; Sottilare, Ragusa, Hoffman, & Goldberg, 2013). In developing their testbed methodology, Hanks et al. reviewed four testbed implementations (Tileworld, the Michigan Intelligent Coordination Experiment [MICE], the Phoenix testbed, and Truckworld) for evaluating the performance of artificially intelligent agents. Although agents have changed substantially in complexity during the past 20–25 years, the methods to evaluate their performance have remained markedly similar.

The ARL adaptive training team designed the GIFT analysis testbed based upon Cohen's assertion (Hanks et al., 1993) that testbeds have three critical roles related to the three phases of research. During the exploratory phase, agent behaviors need to be observed and classified in broad categories. This can be performed in an experimental environment. During the confirmatory phase, the testbed is needed to allow more strict characterizations of agent behavior to test specific hypotheses and compare methodologies. Finally, in order to generalize results, measurement and replication of conditions must be possible. Similarly, the GIFT evaluation methodology (Figure 2) enables the comparison/contrast of ITS elements and assessment of their effect on learning outcomes (e.g., knowledge acquisition, skill acquisition, and retention).

How to Use This Book

This book is organized into four sections:

I. Competency Assessment

II. Evidence-Centered Design and Data Mining

III. General Assessment Methods

IV. Assessment Methods for Particular Domains and Problems

Section I, *Competency Assessment*, describes a variety of assessment methods for modeling long-term proficiency in a particular domain, and how those competency models might be used in GIFT and a variety of learning landscapes. Section II, *Evidence-Centered Design and Data Mining*, highlights the importance of and examines the use of specific evidence to support assessments in ITSs and other instructional systems. Section III, *General Assessment Methods*, discusses the role of assessment methods in ITSs and other instructional systems. Section IV, *Assessment Methods for Particular Domains and Problems*, provides assessment examples within various domains and problem spaces.

Chapter authors in each section were carefully selected for participation in this project based on their expertise in the field as ITS scientists, developers, and practitioners. *Design Recommendations for Intelligent Tutoring Systems: Volume 5 – Assessment Methods* is intended to be a design resource as well as community research resource. Volume 5 can also be of significant benefit as an educational guide for developing ITS scientists, as a roadmap for ITS research opportunities.

References

Aleven, V., McLaren, B., Roll, I. & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education, 16*, 101-128.

Anderson, J. R., Corbett, A. T., Koedinger, K. R. & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*, 167-207.

Anderson, L. W. & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of Educational Objectives: Complete edition*. New York : Longman.

Baker, R.S., D'Mello, S.K., Rodrigo, M.T. & Graesser, A.C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*, 223-241.

Cade, W., Copeland, J. Person, N., & D'Mello, S. K. (2008). Dialogue modes in expert tutoring. In B. Woolf, E. Aimeur, R. Nkambou & S. Lajoie (Eds.), Proceedings of the Ninth International Conference on Intelligent Tutoring Systems (pp. 470-479). Berlin, Heidelberg: Springer-Verlag.

D'Mello, S. & Graesser, A.C. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction, 20*, 147-187.

D'Mello, S. K., Graesser, A. C. & King, B. (2010). Toward spoken human-computer tutorial dialogues. *Human Computer Interaction, 25*, 289-323.

Elson-Cook, M. (1993). Student modeling in intelligent tutoring systems. *Artificial Intelligence Review, 7*, 227-240.

Fletcher, J.D. & Sottilare, R. (2013). Shared Mental Models and Intelligent Tutoring for Teams. In R. Sottilare, A. Graesser, X. Hu, and H. Holden (Eds.) Design *Recommendations for Intelligent Tutoring Systems: Volume I - Learner Modeling*. Army Research Laboratory, Orlando, Florida. ISBN 978-0-9893923-0-3.

Gagne, R. M. (1985). *The conditions of learning and theory of instruction* (4th ed.). New York: Holt, Rinehart & Winston.

Goldberg, B.S., Sottilare, R.A., Brawner, K.W. & Holden, H.K. (2011). Predicting Learner Engagement during Well-Defined and Ill-Defined Computer-Based Intercultural Interactions. In S. D'Mello, A. Graesser, , B.

Schuller & J.-C. Martin (Eds.), *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction (ACII 2011) (Part 1: LNCS 6974)* (pp. 538-547). Berlin Heidelberg: Springer.
Goleman, D. (1995). Emotional intelligence. Bantam Books, New York (1995).
Graesser, A.C., Conley, M. & Olney, A. (2012). Intelligent tutoring systems. In K.R. Harris, S. Graham & T. Urdan (Eds.), *APA Educational Psychology Handbook: Vol. 3. Applications to Learning and Teaching* (pp. 451-473). Washington, DC: American Psychological Association.
Graesser, A. C. & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31,* 104–137.
Hanks, S., Pollack, M.E. & Cohen, P.R. (1993). Benchmarks, test beds, controlled experimentation, and the design of agent architectures. *AI Magazine, 14* (4), 17-42.
Johnson, L. W. & Valente, A. (2008). Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. In M. Goker & K. Haigh (Eds.), Proceedings of the Twentieth Conference on Innovative Applications of Artificial Intelligence (pp. 1632-1639). Menlo Park, CA: AAAI Press.
Krathwohl, D.R., Bloom, B.S. & Masia, B.B. (1964). *Taxonomy of Educational Objectives: Handbook II: Affective Domain*. New York: David McKay Co.
Lepper, M. R., Drake, M. & O'Donnell-Johnson, T. M. (1997). Scaffolding techniques of expert human tutors. In K. Hogan & M. Pressley (Eds), *Scaffolding learner learning: Instructional approaches and issues* (pp. 108-144). New York: Brookline Books.
Litman, D. (2013). Speech and language processing for adaptive training. In P. Durlach & A. Lesgold (Eds.), *Adaptive technologies for training and education*. Cambridge, MA: Cambridge University Press.
Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education*, *10*(1), 98–129.
Murray, T. (2003). An Overview of Intelligent Tutoring System Authoring Tools: Updated analysis of the state of the art. In Murray, T.; Blessing, S.; Ainsworth, S. (Eds.), *Authoring tools for advanced technology learning environments* (pp. 491-545). Berlin: Springer.
Merrill, D., Reiser, B., Ranney, M., & Trafton, J. (1992). Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems. The Journal of the Learning Sciences, 2(3), 277-305
Nkambou, R., Mizoguchi, R. & Bourdeau, J. (2010). *Advances in intelligent tutoring systems*. Heidelberg: Springer.
Patil, A. S. & Abraham, A. (2010). Intelligent and Interactive Web-Based Tutoring System in Engineering Education: Reviews, Perspectives and Development. In F. Xhafa, S. Caballe, A. Abraham, T. Daradoumis & A. Juan Perez (Eds.), *Computational Intelligence for Technology Enhanced Learning. Studies in Computational Intelligence* (Vol 273, pp. 79-97). Berlin: Springer-Verlag.
Person, N. K., Kreuz, R. J., Zwaan, R. A. & Graesser, A. C. (1995). Pragmatics and pedagogy: Conversational rules and politeness strategies may inhibit effective tutoring. *Cognition and Instruction, 13*(2), 161–188.
Psotka, J. & Mutter, S.A. (1988). *Intelligent Tutoring Systems: Lessons Learned*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Rus, V. & Graesser, A.C. (Eds.) (2009). The Question Generation Shared Task and Evaluation Challenge. Retrieved from http://www.questiongeneration.org/.
Simpson, E. (1972). The classification of educational objectives in the psychomotor domain: *The psychomotor domain*. Vol. 3. Washington, DC: Gryphon House.
Sleeman D. & J. S. Brown (Eds.) (1982). *Intelligent Tutoring Systems*. Orlando, Florida: Academic Press, Inc.
Soller, A. (2001). Supporting social interaction in an intelligent collaborative learning system. *International Journal of Artificial Intelligence in Education, 12*(1), 40-62.
Sottilare, R. & Gilbert, S. (2011). Considerations for tutoring, cognitive modeling, authoring and interaction design in serious games. *Authoring Simulation and Game-based Intelligent Tutoring workshop at the Artificial Intelligence in Education Conference (AIED) 2011*, Auckland, New Zealand, June 2011.
Sottilare, R., Holden, H., Brawner, K. & Goldberg, B. (2011). Challenges and Emerging Concepts in the Development of Adaptive, Computer-based Tutoring Systems for Team Training. *Interservice/Industry Training Systems & Education Conference*, Orlando, Florida, December 2011.
Sottilare, R.A., Brawner, K.W., Goldberg, B.S. & Holden, H.K. (2012). *The Generalized Intelligent Framework for Tutoring (GIFT)*. Orlando, FL: U.S. Army Research Laboratory Human Research & Engineering Directorate (ARL-HRED).
Sottilare, R. (2012). Considerations in the development of an ontology for a Generalized Intelligent Framework for Tutoring. *International Defense & Homeland Security Simulation Workshop* in Proceedings of the I3M Conference. Vienna, Austria, September 2012.

Sottilare, R. (2013). Special Report: Adaptive Intelligent Tutoring System (ITS) Research in Support of the Army Learning Model- Research Outline. *Army Research Laboratory* (ARL-SR-0284), December 2013.

Sottilare, R., Ragusa, C., Hoffman, M. & Goldberg, B. (2013). Characterizing an adaptive tutoring learning effect chain for individual and team tutoring. In Proceedings of the *Interservice/Industry Training Simulation & Education Conference*, Orlando, Florida, December 2013.

Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017. DOI: 10.13140/RG.2.2.12941.54244.

Sottilare, R.A., Burke, C.S., Salas, E., Sinatra, A.M., Johnston, J.H. & Gilbert, S.B. (2017). Towards a Design Process for Adaptive Instruction of Teams: A Meta-Analysis. *International Journal of Artificial Intelligence in Education.* DOI: 10.1007/s40593-017-0146-z.

VanLehn, K. (2006) The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education. 16*(3), 227-265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist, 46*(4), 197-221.

Woolf, B.P. (2009). *Building intelligent interactive tutors*. Burlington, MA: Morgan Kaufmann Publishers.

SECTION I
COMPETENCY ASSESSMENT

Dr. Robert Sottilare, Ed.

CHAPTER 1 – Understanding Competency Assessment Methodologies Applied to Adaptive Instruction in Intelligent Tutoring Systems

Robert Sottilare, Ph.D.
US Army Research Laboratory

Introduction

The first goal of this chapter is to provide the reader with a basic understanding of competence, competency assessment methods, and their relationship to processes leading to knowledge and skill acquisition during adaptive instruction. The second goal is to introduce the chapters in this section of the book and define each chapter's importance to competency assessment and the design of Intelligent Tutoring Systems (ITSs). Let's begin by defining competence and competency. In terms of learning, competence is an "ability or skill" or "the ability to do something successfully or efficiently" (Merriam-Webster, 2017). The goal of adaptive instruction is to guide the learner in developing competence in one or more domains.

Adaptive instruction delivers content, offers feedback, and intervenes with learners based on tailored strategies and tactics with the goal of optimizing learning, performance, retention, and transfer of skills for both individual learners and teams. The Generalized Intelligent Framework for Tutoring (GIFT; Sottilare, Brawner, Sinatra & Johnston, 2017) is a tutoring architecture that has evolved over the last five years with three primary goals: 1) reduce the time and skill required to author ITSs, 2) automate best practices of instruction in the policy, strategies, and tactics of tutoring, and 3) provide a testbed to assess the effectiveness of adaptive instructional tools and methods with respect to learning, performance, retention, and transfer of skills. The cycle of adaptive instruction includes 1) assessment of the learner's states (e.g., performance), 2) selection of tailored strategies (plans for action) and application of tactics (actions) by the tutor, and 3) evaluation of the effect of tutor strategies/tactics on learning and performance.

According to Person, Graesser, Kreuz & Pomeroy (2003), the tutoring process consists of 5 steps to determine knowledge or skill. During the first three steps in a dialogue process between the learner and the tutor: 1) the tutor asks a question, 2) the learner answers it, and 3) feedback is received from the tutor about the learner's answer. In step 4, the tutor and the learner work together to improve the answer collaboratively, and finally in step 5, the tutor assesses the learner's understanding of their answer and this result provides an initial indication of competence. By following this tutoring process, the ITS is harnessing the power of previous approaches that have been shown to be successful with human tutors.

This is an important part of the learning process and while competence can be demonstrated during a single performance, the idea that a learner demonstrates competency in a complex domain really requires more evidence. The ability of the learner to demonstrate knowledge and skill over time and in a variety of situations demonstrates a higher degree of confidence in their competency than just a single event. So, competency is more of a long-term assessment of skills and abilities in a domain rather than just a single measure of ability. The discussions that follow provide a short description of each chapter in the "competency" section of this book and it importance in relationship to ITS design.

Competence-based Knowledge Structures and Current Challenges for E-Assessment

Our first chapter in this section by Dietrich Albert, Alexander Nussbaumer, Bor-Chen Kuo, Peter W. Foltz, and Xiangen Hu presents challenges for electronic assessment. Electronic assessment (also known as e-assessment, online assessment, computer-assisted assessment, computer-mediated assessment, or computer-based assessment) is the use of information technology in various forms of assessment, but in this case is focused on instructional assessment of competency in training and educational domains. The chapter presents an overview of Competence-based Knowledge Space Theory (CbKST) for knowledge and competence assessment. According to CbKST, the knowledge and skills in a domain are a known set with the learner familiar with some subset of this domain set. This approach allows ITS developers to set standards and goals for development based on the hierarchical relationship of knowledge and skills in a given domain and thereby lends structure to both the authoring and assessment processes for ITSs. CbKST also supports personalization of the tutoring experience based on competency assessment or a mapping of what the learner knows, what they should learn next, and which concepts should be reviewed soon.

Also discussed in this chapter are new psycho-pedagogical concepts that imply the need for new assessment approaches. These concepts include open-learner models (OLMs), learning analytics, game-based learning, self-regulated learning (SRL), and various learning environments. The mapping of competencies in OLMs provide the learner, peers, and instructors the opportunity to understand where the learner needs to improve. In ITSs, a standardized visualization or mapping strategy would also for e-assessment of domain competency, which could be used for tailoring future tutoring experiences. Learning analytics include methods to acquire, analyze, and report information about the learner that influences their learning and performance. Understanding what in human variability contributes to learning allows designers of ITSs to track and respond to learning factors in real time during tutoring.

Games are engaging experiences, which when combined with learning content and measures of assessment, make effective learning tools. Games allow ITS designers to motivate learners and unobtrusively assess their performance while also enticing them to compete against other and themselves through badging. SRL is the degree to which learners are "metacognitively, motivationally, and behaviorally active participants in their own learning process". Learning goals can vary in time and intensity. Allowing learners to control learning goals and select learning experiences empowers them and increases the probability of return engagements with computer-based tutors in the future. As learners become more familiar and confident with ITS technology and more competent in a domain of study, more effort should be made to design ITSs to offer more control to learners. Finally, new learning environments (e.g., learning management systems [LMSs] and online courses) require the development of competence with the use of the tools and methods that are part of that learning environment. Care should be taken to treat domain assessments in new learning environments with a lower degree of confidence until the learner has demonstrated tool competency.

Total Learning Architecture (TLA)

The next two chapters in section focus on a broad learning landscape of providers and consumers called the Total Learning Architecture (TLA) and its measure of competence, the experience application program interface (xAPI). The TLA is a vision of the next-generation integrated learning environment. The second chapter by Gregory Goodwin, J.T. Folsom-Kovarik, Andy Johnson, Sae Schatz, and Robert Sottilare is focused on how learning experience providers like GIFT might be integrated within the TLA and contribute to domain competence assessment. Potential methods for experience tracking, competency management assessment, learner modeling, and content brokering between GIFT and the TLA are discussed along with recommended standards for data exchange. Experience tracking is based on a series of xAPI achievement

statements that might be provided as at various levels of granularity (e.g., completed a degree, completed a course, completed an assignment, or completed a problem). The specifics of how these xAPI achievement statements might be weighed to provide a clear picture of domain competency is a challenge that hasn't been completely worked out yet, but the building blocks for a competency model are available.

The third chapter by Andy Johnson, Benjamin D Nye, Diego Zapata-Rivera, and Xiangen Hu focuses on GIFT and xAPI alignment in five main areas: 1) fine-grained achievement data, such as answering specific questions, 2) calculating duration to provide better metrics when compared with learner outcomes 3) mechanisms to allow a learner's to assess the quality of their experience, 4) an effective model of competency and learning/forgetting, and 5) an assessment profile within xAPI to track both formal and informal learning experiences.

Big Data, Career Management, and Competencies

The fourth chapter in the competency section of this book by Brent Olde discusses the role of data analytics in career management and competency development. Olde puts forth a vision for career management in the US Navy that uses data analytic techniques to assess 1) competence and readiness from the individual to the collective level, 2) training efficiencies and effectiveness, 3) optimal personnel assignments, 4) transfer of training, and 5) training requirements based on current operational needs. Modeling skill needs, skill development, and skill transfer could provide organizational decision makers with a tool to shape both organizations and personnel to meet shared goals.

Coordinating Evidence across Learning Modules Using Digital Badges

The fifth chapter in our competency section by Ross Higashi, Christian Schunn, Vu Nguyen, and Scott Ososky discusses the use of digital badges as a mechanism for capturing evidence of domain competency across a variety of learning experiences. Much like the xAPI standard writes out achievement statements to a long-term learner record, badges could be issued as evidence of learning across different learning experiences and sources. The authors envision badges as a method for selection and projection of their success: 1) assessing competency to be a successful in a particular job or environment (college) based on prior experiences, or 2) selecting a next task for learning based on prior experiences. Toward this end, they provide three design guidelines for a badge-based, evidence framework: 1) the framework must support the inclusion of multiple kinds of evidence concurrently, 2) the design must recognize and represent multiple "levels" of evidence and define how much evidence is sufficient, and 3) each type of evidence in the badge must be summarizable for quick viewing, and the composite strength of the badged evidence claim should also be easily summarized. Badges have the advantage of being simple and easy to interpret whereas xAPI statements provide direct evidence of learning, they are more difficult to summarize.

Leveraging Domain Models for Personalizing Problem Solving and Learning

The sixth chapter in this section by Louise Yarnall, Eric Snow, Erica Snow, and Irvin Katz examines the basic elements of evidence-centered design (ECD) and describes the application of ECD to the complex skills of computational thinking and science inquiry. The chapter also discusses the implications for content authoring and assessment of problem solving tasks in a GIFT-based tutor. The authors assert that ECD modeling can support standardized approaches in the design of assessments of the most complex forms of applied reasoning and problem solving. The benefits of ECD modeling focuses on its theoretical soundness, consistency of documentation format, and potential for model adaptation across multiple domains and are well suited for a multi-domain framework like GIFT.

Assessing Individual Learner Performance in MOOCs

The final chapter in this section by Ryan Baker, Piotr Mitros, Benjamin Goldberg, and Robert Sottilare examines assessment of individual learner performance during massive open online courses (MOOCs). This chapter describes an ongoing effort between ARL, Carnegie Mellon and University of Pennsylvania to integrate GIFT with LMS sites like edX. This effort expands the domains in which GIFT can guide learning, but also presents some challenges in assessment.

The initial phase of the project developed a GIFT interface with the learning tools interoperability (LTI) component of edX. This enabled MOOC developers to reference GIFT-managed lessons within their structure and to receive data back following the completion of a GIFT lesson for performance tracking and accreditation purposes. With the LTI component in place, the next phase will involve configuring MOOC content into a set of lessons that adhere to the authoring standards and run-time schemas of GIFT.

References

Merriam-Webster Online Dictionary. (2017). Competency. https://www.merriam-webster.com/dictionary/competency.

Person, N. K., Graesser, A. C., Kreuz, R. J. & Pomeroy, V. (2003). Simulating human tutor dialog moves in AutoTutor. *International Journal of Artificial Intelligence in Education (IJAIED)*, *12*, 23–39.

Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017. DOI: 10.13140/RG.2.2.12941.54244.

CHAPTER 2 – Competence-based Knowledge Structures and Current Challenges for E-Assessment

Dietrich Albert[1,2], Alexander Nussbaumer[1], Bor-Chen Kuo[3], Peter W. Foltz[4], and Xiangen Hu[5]
Graz University of Technology[1], University of Graz[2] , National Taichung University of Education[3], University of Colorado Boulder & Pearson[4], University of Memphis[5]

This chapter presents a selection of new challenges for e-assessment. The first part presents a short overview of Competence-based Knowledge Space Theory (CbKST) and related methods for knowledge and competence assessment. Based on CbKST, the second part presents new psycho-pedagogical concepts and technologies that also imply the need for new assessment approaches. The psycho-pedagogical concepts include self-regulated learning and metacognition, as well as a need for non-invasive assessment that does not disturb the learner. The new technologies include open learner models and visualization, new learning environments from individually composed environments to augmented reality. Several of these new assessment approaches include the need for interpreting the learners behavior in terms of competences in different environments, for example, the interactions with computer systems and the real world.

Introduction

Technology-enhanced learning (TEL) and distance learning has been playing an important role in psycho-pedagogical research and educational practice in the last decades. Intelligent tutoring systems (ITSs) and adaptive systems played a major role in research, because they aimed at tailoring the learning content and the learning environment to the pre-knowledge, learning progress competences, preferences, needs, and goals of the individual learners (Albert & Schrepp, 1999). To adapt the content and system's behavior, data and information about the learner has to be known and stored in a structured way as a user or learner model. E-assessment of the learner's knowledge and so on is a way to gain this information and enables an adaptive system to personalize the content and the learning environment.

This chapter focuses on challenges for e-assessment and examines approaches from the point of view of Competence-based Knowledge Space Theory (CbKST). CbKST is a prominent framework for structuring knowledge, adaptive assessment, and personalization of the learning experience. It is based on a psychological-mathematical framework that allows to translate this method into a system design and development. CbKST has been used in commercial as well as in research applications to structure knowledge domains, define competences in such knowledge domains, provide assessment procedures, and adapt the learning content and learning trajectories. Different models and algorithms have been used for these purposes (e.g., Falmagne et al., 2013).

While CbKST and e-assessment methods related to it could be applied seamlessly on adaptive systems in the past, new concepts and advancements in technology-enhanced learning require new ways of integration. Recent developments in the psycho-pedagogical field are evident and have strong effect on the methods, requirements, and role of e-assessment. For instance, self-regulated learning approaches aim to empower the learner to take over the control of their own learning process instead of letting the system decide on the learning alternatives. Game-based learning intends to integrate fun to play and learning by stimulating a flow experience. Moreover, new test-theoretical algorithms and methods were elaborated that allow for real-time assessment and adaptation.

In addition to the psycho-pedagogical field, new technical developments in the context of technology-enhanced learning took place during the last years. For example, learning analytics methods are used to analyze the learning behavior and provide feedback to the learner and the teacher. Semantic technologies are used to create meaningful models and structures of text-based content and student knowledge. Virtual reality applications emerge from new hardware improvements, which allow the use of a new type learning environment.

These new developments lead to new situations regarding e-assessment opportunities and procedures. Such new situations require the modernization of assessment procedures to fully exploit the potential of these innovations. This chapter investigates the aforementioned psycho-pedagogical and technical innovations in the light of assessment and e-assessment. The new opportunities, as well as new challenges and requirements, are elaborated and explained.

Competence-based Knowledge Space Theory and Assessment

Knowledge Space Theory

Knowledge Space Theory (KST) is a mathematical-psychological theory for representing domain and learner knowledge (Doignon & Falmagne, 1985, 1999; Falmagne & Doignon, 2011; Albert & Lukas, 1999; Falmagne et al., 2013). In KST, a knowledge domain is identified with a set Q of problems. The subset of problems that a person is able to solve represents the knowledge state of this individual. Among the problems of a domain mutual dependencies will exist, such that not all potential knowledge states (i.e., subsets of problems) will actually occur. In KST's simplest version, these dependencies are captured by a so-called prerequisite relation (also referred to as precedence relation or surmise relation), which restricts the number of possible knowledge states. Two problems, a and b, are in a prerequisite relation whenever the correct solution of problem a is a prerequisite for the mastery of problem b. Illustrated in a Hasse diagram (Figure 1), ascending sequences of line segments indicate a prerequisite relationship. The collection of knowledge states corresponding to a prerequisite relation is called a knowledge structure. In a knowledge structure, a range of different learning paths from the naive knowledge state to the expert knowledge state are possible (as shown in Figure 1).

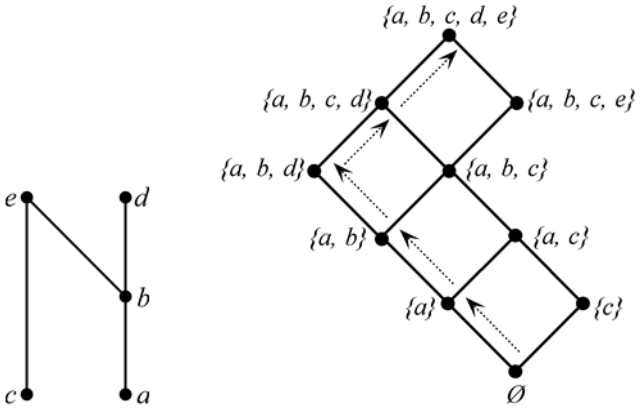


Figure 1. Example of a prerequisite relation and the induced knowledge structure. Dashed arrows show a possible learning path (from Albert et al., 2012, p. 26).

A knowledge structure enables adaptive assessment procedures for efficiently identifying the current knowledge state of an individual (see, e.g., Doignon & Falmagne, 1999; Hockemeyer, 2002). Through defining individual starting and goal states for a learner, meaningful learning sequences with reasonable choices for navigation and appropriate levels of challenge can be realized for each learner.

The commercial Adaptive Learning with Knowledge Spaces (ALEKS; http://www.aleks.com) system is a fully automated, multilingual, adaptive tutor that grounds on KST (Canfield, 2001). The system provides individualized learning including explanations, practice, and feedback on learning progress for various disciplines, especially for mathematics, chemistry, and business administration. ALEKS adaptively and accurately assesses which concepts a learner already knows, what the learner is ready to learn next, which previously learned material should be addressed for review, and continuously updates a precise map of the learner's knowledge state (Falmagne, Doignon, Cosyn & Thiery, 2004; Hardy, 2004; Falmagne et al., 2013)

Competence-based Knowledge Space Theory

CbKST incorporates psychological assumptions on underlying skills and competencies that are required for solving the problems under consideration (Düntsch & Gediga, 1995; Korossy, 1997, 1999; Heller et al., 2006, 2013a,b). This approach assigns to each problem a collection of skills that are needed to solve this problem and to learning objects those skills that they teach. Similar to the knowledge state a competence state can be defined that consists of a set of skills that the learner has available. Furthermore, there may also be a prerequisite relationship between skills.

CbKST provides algorithms for efficient adaptive assessment to determine the learner's current knowledge and competence state, which builds the basis for personalization purposes. Based on this learner information, personalized learning paths can be created. Goal setting can be done by defining skills to be achieved (competence goal) or problems to be capable of solving. The competence gap to be closed during learning is represented by the skills that are part of the goal but not part of the competence state of a learner.

Several implementation approaches make use of CbKST. For instance, the concept of competence learning structures were applied in the prototypical adaptive learning system Adaptive & Personalized E-Learning System (APeLS; Conlan, Hockemeyer, Wade & Albert, 2002). It can easily merge content from different sources to build an adaptive course. The only requirement is that the individual learning objects carry metadata information on required and taught competencies according to the competence learning structure approach given that the metadata author use the same competence terminology.

Current Challenges in E-Assessment

The previous section gave an overview of well-known assessment methods in the field of CbKST. However, in recent years, new psycho-pedagogical concepts and new technological developments have been observed that have influence on e-assessment. This section focuses on the current potential for e-assessment offered by CbKST.

Open-Learner Models (OLMs)

Learner models are a core component of ITSs and adaptive systems that use them for their internal strategies (Brusilovsky et al., 2007). The result of an assessment is typically a set of solved assessment items, available competences, or a probability distribution regarding knowledge and competences. In TEL systems, such results are typically stored in leaner models and used for adaptation and personalization.

A newer trend follows the idea of opening up these models to the user to support the learners in their reflection of the learning process and support teachers to better understand their students. A promising opportunity how to facilitate reflection and how to raise awareness is represented by OLMs, where visualization, trust, and credibility play the key roles (Bull & Kay, 2010). OLMs provide suitable interfaces for users to enable them to view and, in some cases, also to change their learner model. This information can be made available also to others (peers and teachers), who can assist learning of the user. Examples of OLM visualization techniques are overview-zoom-filter approaches including tree maps, tag clouds, or sunburst views (Bull & Kay, 2010; Mathews et al., 2012).

Learning Analytics (LA)

The Society for Learning Analytics Research (http://solaresearch.org/) defines LA as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs". In the New Media Consortium (NCM) Horizon Report 2014, LA is described as a rapidly developing trend in higher education, where learning is happening more and more within online and hybrid environments (Johnson et al., 2014). According to this report, LA can potentially help to transform education from a standard one-size-fits-all approach into responsive and flexible frameworks.

In this way, LA approaches offer new possibilities for e-assessment. Instead of conducting explicit assessment at certain points during the learning process, LA approaches offer the possibility of non-invasive assessment by collecting and analyzing interaction data of learners. When learners interact with a learning environment and make use of its content and functionalities, data about this interaction can be collected and stored. An analytics module can analyze these data regarding knowledge and competences a learner might have. This approach requires an appropriate method how interaction data can be analyzed so that it results in a trustworthy assessment. This is even more important if the collected data are unstructured, from different modules, or from an incoherent system.

Game-Based Learning

Digital learning games represent an e-learning technology that is increasingly recognized by educational practitioners (Johnson et al., 2014). With their highly engaging and motivating characters, games constitute effective educational tools for creating authentic learning tasks and meaningful, situated learning (De Freitas, 2013). An important characteristic of serious games is the flow concept (Csikszentmihalyi & LeFevre, 1989), which describes a situation when people are highly engaged and lose track of time, which often happens, when competence development and game challenges are balanced (e.g., the game should not demand too much or too little from the player). This concept indicates that the games can make use of the learners' competences and thus they need to get this information.

In order not to disturb the flow experience, the assessment has to be done in a non-invasive way and in real time. Besides LA approaches (as described previously), the concept of micro-adaptivity (Stefanutti & Albert, 2003; Albert et al. 2007; Kickmeier-Rust et al., 2008) has been employed to assess the learners' competences during game play. Based on a model of competences, tasks, and learner activities, the learner's behavioral actions can be interpreted toward a competences assessment. The non-invasive updating of the to be assessed competences during complex problem solving or game-based learning in real time requires very fast algorithms and calculations (Augustin et al., 2011, 2013, 2015).

Self-Regulated Learning (SRL)

According to Zimmerman (2002) students can be described as self-regulated to the degree that they are metacognitively, motivationally, and behaviorally active participants in their own learning process. To define students' learning as self-regulated, they have to use specific strategies for attaining their goals and their learning behavior has to be based on self-efficacy perceptions. In self-regulated learning the learners are active and able to control, monitor, and regulate their cognition, motivational state, behavior, and context. Furthermore, the learners set goals and try to achieve them through progress-monitoring. This type of learning is especially important when using digital learning environments because no teacher or tutor guides them.

Assessment of SRL competences is significantly more difficult than the assessment of domain knowledge and almost not done yet. In principle, three approaches are known. The first is self-assessment through the use of SRL questionnaires (e.g., Fill Giordano et al., 2010). Such questionnaires asks questions regarding the own SRL capabilities where the items are related to different SRL competences. The second is asking teachers or tutors by employing a questionnaire where teachers can assess the SRL competences of their students (Mikroyannidis et al., 2013). Third, a learning analytics approach can be used that tracks the learners' behavior and relates it to SRL competences by using underlying models and analytics methods (Nussbaumer et al., 2012).

An important phase in SRL is self-reflection which can be fostered by experiencing the gap between an objective assessment and a subjective self-evaluation/assessment. Respective tools for supporting these assessments have been developed and provided by Nussbaumer (2008), for example.

Learning Environments

Traditionally, technology for learning support was centered on learning management systems (LMSs). They primarily focus on distributing learning content, organizing the learning processes, and serving as interface between learner and teacher. In educational institutions LMSs have become very popular and are used in many universities and schools (Paulsen, 2003). Examples of LMSs are Moodle, CLIX, Blackboard, WebCT, Sakai, ILIAS and .LRN. They all have in common that different tools are integrated in a single system, such as discussion forums, file sharing, whiteboards, chat, and e-portfolios (Dalsgaard, 2006). These tools together with learning content are bundled by teachers or tutors, which leads to a centralized and standardized learning experience.

In the last years, many learning environments were developed that are fundamentally different from the traditional learning management systems. Personal learning environments (PLEs) strive for a more natural and learner-centric approach and is characterized by the freedom that individual learners have to select and control services and tools they use. Virtual reality (VR) environments allow full immersion in a virtual world where the learners can freely navigate. Augmented reality (AR) solutions allow the overlay of digital information over the physical world. Typically, the AR applications recognize the physical world and the behavior of the learner in the physical world, and provide feedback and hints.

These new types of digital environments entail new challenges for e-assessment. PLEs are individually composed and require a flexible design for assessment, when and how it should happen. VR applications require the interpretation of the learners' behavior. Complex virtual worlds allow a magnitude of different actions, which makes the assessment difficult. Assessment with AR applications is even more difficult, since it relies on the interpretation of real-world behavior.

To assess the learners' ability in composing and handling their PLEs, different types of competences have to be taken into account – like the competences to select (Berthold et al,. 2011) and use the selected tools (tool competence). Respective behavioral data have to be collected, analyzed and interpreted. A psycho-pedagogical model developed for SRL in PLEs may become the basis to assess the different types of competences by interpreting the learner's actions (Nussbaumer et. al, 2014).

Conclusion and Outlook

This chapter presented a selection of new challenges for e-assessment. The first part presented an overview of CbKST and related methods for knowledge and competence assessment. The second part presented new psycho-pedagogical concepts and technologies that also imply the need for new assessment approaches. Many of these approaches include the interpretation of the learners behavior in different environments, for example, the interactions with computer systems and the real world.

Beside the presented challenges for e-assessment, there are further topics to be considered in the future. For example, recognizing and assessing misconceptions would be beneficial in e-learning. Detecting key competences can bring significant benefit for the learner. The use of open content is a challenge for e-assessment, because the assessment methodology has to be adapted automatically to the visited content. While traditional assessment focuses on content, new methods of assessment also should take into account meta-cognitive skills, motivation, preferences, social contexts, and many other learner characteristics. Furthermore, in open environments (in terms of content, peers, and tools) privacy and data protection aspects have to be respected.

References

Albert, D., Hockemeyer, C., Conlan, O. & Wade, V. (2001). Reusing adaptive learning resources. In C.-H. Lee et al., editor, Proceedings of the International Conference on Computers in Education ICCE/SchoolNet2001, volume 1, pp. 205–210.

Albert, D., Hockemeyer, C., Kickmeier-Rust, M. D., Peirce, N. & Conlan, O. (2007). Microadaptivity within complex learning situations – a personalized approach based on competence structures and problem spaces. In Chen, W. & Ogata, H. (Eds) learning by effective utilization of technologies: facilitating intercultural understanding Supplementary Proceedings of the 15th International Conference on Computers in Education (ICCE 2007) Hiroshima, Japan. https://www.scss.tcd.ie/owen.conlan/publications/icce2007_conlan.pdf.

Albert, D., Hockemeyer, C., Kickmeier-Rust, M. D., Nussbaumer, A. & Steiner, C. M. (2012). E-learning based on metadata, ontologies and competence-based knowledge space theory. In D. Lukose, A. R. Ahmad & A. Suliman (Eds.), Knowledge technology: Third knowledge technology week, KTW 2011, Kajang, Malaysia, July 18-22, 2011. Revised selected papers (pp. 24–36). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-32826-8_3.

Albert, D. & Lukas, J. (Eds.). (1999). Knowledge spaces: Theories, empirical research, and applications. Mahwah, NJ: Lawrence Erlbaum Associates.

Albert D, Schrepp M. (1999). Structures and design of an intelligent tutoring system based on skill assignments. In: Albert D, Lukas J, editors. Knowledge spaces: theories, empirical research, and applications. Mahwah: Lawrence Erlbaum; p. 179–96.

Augustin, T., Hockemeyer, C., Kickmeier-Rust, M. D. & Albert, D. (2011). Individualized skill assessment in digital learning games: Basic definitions and mathematical formalism. IEEE Transactions on Learning Technologies, 4(2), 138–148. doi:10.1109/TLT.2010.21.

Augustin, T., Hockemeyer, C., Kickmeier-Rust, M. D., Podbregar, P., Suck, R. & Albert, D. (2013). The simplified updating rule in the formalization of digital educational games. Journal of Computational Science, 4(4), 293–303. doi:10.1016/j.jocs.2012.08.020.

Augustin, T., Hockemeyer, C., Suck, R., Podbregar, P., Kickmeier-Rust, M. D. & Albert, D. (2015). Individualized skill assessment in educational games: The mathematical foundations of partitioning. Journal of Mathematical Psychology, 67, 1–7. doi:10.1016/j.jmp.2015.05.003.

Berthold, M., Pachtchenko, S., Kiefel, A., Nussbaumer, A. & Albert, D. (2011). Identifying requirements for a psycho-pedagogical mash-up design for personalising the learning environment. PALE-2011 Personalization Approaches in Learning Environments: Proceedings of the International Workshop on Personalization Approaches in Learning Environments, Girona, Spain. 36–40.

Brusilovsky, P. & Millán, E. (2007). User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: P. Brusilovsky, A. Kobsa & W. Nejdl (eds.): *Lecture Notes in Computer Science* 4321 (The Adaptive Web). Springer Berlin-Heidelberg, pp. 3–53.

Bull, S. & Kay, J. (2010). *Open learner models. In: Advances in Intelligent Tutoring Systems*. Springer Berlin-Heidelberg, pp. 301–322.

Canfield, W. (2001). ALEKS: a Web-based intelligent tutoring system. *Mathematics and Computer Education*, 35 (2), 152–158.

Csikszentmihalyi, M., and LeFevre, J. (1989). Optimal experience in work and leisure. *J. Pers. Soc. Psychol*. 56, 815–822. doi: 10.1037/0022-3514.56.5.815.

Dalsgaard, C. (2006). Social Software: E-learning Beyond Learning Management Systems. European. Journal of Open, Distance and E-learning 2006(2).

De Freitas, S. (2013). Learning in immersive worlds. A review of game-based learning. JISC E-learning programme. Retrieved March 1, 2013 from http://www.jisc.ac.uk/media/documents/programmes/elearninginnovation/gamingreport_v3.pdf.

Doignon, J. & Falmagne, J. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, 23, 175–196.

Doignon, J. & Falmagne, J. (1999). *Knowledge Spaces*. Berlin: Springer.

Düntsch, I. & Gediga, G. (1995). Skills and knowledge structures. *British Journal of Mathematical and Statistical Psychology*, 48, 9–27.

Falmagne, J.-C., Albert, D., Doble, C., Eppstein, D. & Hu, X. (Eds.). (2013). Knowledge spaces: Applications in education. Berlin Heidelberg: Springer.

Falmagne, J.-C. & Doignon, J.-P. (2011). Learning spaces. Interdisciplinary applied mathematics. Berlin: Springer.

Falmagne, J.-C., Doignon, J.-P., Cosyn, E. & Thiery, N. (2004). The assessment of knowledge in theory and practice. Retrieved August 23, 2011 from http://www.aleks.com/about/Science_Behind_ALEKS.pdf.

Fill Giordano, R., Litzenberger, M. & Berthold, M. (2010). On the Assessment of strategies in self-regulated learning (SRL)–differences in adolescents of different age group and school type (p. Poster). Salzburg: 9. Tagung der Österreichischen Gesellschaft für Psychologie, Salzburg.

Hardy, M.E. (2004). Use and evaluation of the ALEKS interactive tutoring system. *Journal of Computing Sciences in Colleges*, 19 (4), 342–347.

Heller, J., Augustin, T., Hockemeyer, C., Stefanutti, L. & Albert, D. (2013a). Recent developments in competence-based knowledge space theory. In J. Falmagne, D. Albert, C. Doble, D. Eppstein & X. Hu (Eds.), Knowledge spaces: Applications in education (pp. 243-286). Berlin, Heidelberg: Springer Berlin Heidelberg.

Heller, J., Steiner, C., Hockemeyer, C. & Albert, D. (2006). Competence-based knowledge structures for personalised learning. *International Journal on E-Learning*, 5(1), 75–88.

Heller, J., Ünlü, A. & Albert, D. (2013b). Skills, competencies and knowledge structures. In J. Falmagne, D. Albert, C. Doble, D. Eppstein & X. Hu (Eds.), Knowledge spaces: Applications in education (pp. 229–242). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-35329-1_11.

Hockemeyer, C. (2002). A Comparison of Non–Deterministic Procedures for the Adaptive Assessment of Knowledge. *PsychologischeBeiträge*, 44, 495–503.

Hockemeyer, C., Conlan, O., Wade, V. & Albert, D. (2003). Applying competence prerequisite structures for eLearning and skill management. Journal of Universal Computer Science, 9, 1428–1436.

Johnson, L., Adams Becker, S., Estrada, V. & Freeman, A. (2014). NMC Horizon Report 2014: Higher Education Edition. The New Media Consortium. Retrieved from http://www.nmc.org/publications/2014-horizon-report-higher-ed.

Kickmeier-Rust M. D., Hockemeyer, C., Albert, D. & Augustin T.(2008). Micro Adaptive, Non-invasive Knowledge Assessment in Educational Games. Second IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning, Banff, BC, 2008, pp. 135-137. doi: 10.1109/DIGITEL.2008.10.

Korossy, K. (1997). Extending the Theory of Knowledge Spaces: A Competence-Performance Approach. Zeitschrift für Psychologie, 205, pp. 53–82.

Korossy, K. (1999). Modeling Knowledge as Competence and Performance. In D. Albert & J. Lukas (Eds.), Knowledge Spaces: Theories, Empirical Research Applications (pp. 103–132). Mahwah, NJ: Lawrence Erlbaum Associates.

Mathews, M., Mitrovic, A., Lin, B., Holland, J. & Churcher, N. (2012). Do Your Eyes Give it Away? Using Eye-Tracking Data to Understand Students' Attitudes Towards Open Student Model Representations, In S.A. Cerri, W.J. Clancey, G. Papadourakis and K. Panourgia (Eds), Intelligent Tutoring Systems (pp. 422–427), Berlin-Heidelberg: Springer-Verlag.

Mikroyannidis, A., Connolly, T., Law, E., Schmitz, H.-C., Vieritz, H., Nussbaumer, A., Berthold, M., Ullrich, C. & Dhir, A. (2014). Self-Regulated Learning in Formal Education: Perceptions, Challenges and Opportunities. International Journal of Technology Enhanced Learning, 6(2), 145–163.

Nussbaumer, A. (2008). Supporting Self-Reflection through Presenting Visual Feedback of Adaptive Assessment and Self-Evaluation Tools. Proceedings of the 11th International Conference on Interactive Computer-aided Learning (ICL 2008).

Nussbaumer, A., Kravcik, M., Renzel, D., Klamma, R., Berthold, M. & Albert, D. (2014). A Framework for Facilitating Self-Regulation in Responsive Open Learning Environments. Retrieved from http://arxiv.org/abs/1407.5891.

Nussbaumer, A., Scheffel, M., Niemann, K., Kravcik, M. & Albert, D. (2012). Detecting and Reflecting Learning Activities in Personal Learning Environments. Proc. of the 2nd Workshop on Awareness and Reflection in Technology-Enhanced Learning (artel12) at European Conference for Technology-Enhanced Learning 2012 (EC-TEL). Saarbrucken.

Paulsen, M (2003). Experiences with Learning Management Systems in 113 European Institutions. Educational Technology & Society 6(4), 134–148.

Stefanutti, L. & Albert, D. (2003). Skill assessment in problem solving and simulated learning environments. Journal of Universal Computer Science, 9(12), 1455–1468. doi:10.3217/jucs-009-12-1455.

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. Theory Into Practice, 41(2), 64–70. doi: 10.1207/s15430421tip4102_2.

CHAPTER 3 – Exploring Assessment Mechanisms in the Total Learning Architecture (TLA)

Gregory Goodwin[1], J. T. Folsom-Kovarik[2], Andy Johnson[3], Sae Schatz[3], and Robert Sottilare[1]
US Army Research Laboratory[1], Soartech[2], Advanced Distributive Learning Initiative[3]

Introduction

The focus of this chapter is on the challenges and potential solutions to conducting real-time and long-term assessments of performance, learning, and domain competency in the Total Learning Architecture (TLA). TLA, a distributed learning ecosystem, is being developed by the US Office of the Secretary of Defense to support capabilities for instruction anytime and anywhere. TLA is an evolving set of standardized specifications that enable responsible sharing of essential learning data between applications using common interfaces and data models. The applications that could be part of the TLA ecosystem range from simple desktop applications to immersive simulations to mobile apps, and would serve as either service providers or consumers. Expected services include applications like intelligent tutoring systems (ITSs; e.g., AutoTutor, Cognitive Tutor, or Generalized Intelligent Framework for Tutoring [GIFT]-based tutor), which provide information to other services and consume information from other services.

The TLA is expected to provide services including experience tracking, competency assessment, learner modeling, and content brokering. All of these fundamentally involve learner assessments. Experience tracking (via the experience application programming interface [xAPI]) provides a standard for encoding and storing data about learners' interactions with learning experiences and applications, providing fine-grained evidence that can make assessment precise and timely. TLA will also establish a common way for systems to reference and represent competencies and competency relationships, supporting assessment sharing. Learner models will contain data about assessed mastery of competencies as well as traits, preferences, individual differences, and demographic data. Learner models that are broadly accessible to learning applications will support up-to-date and accurate competency assessment. Content brokering (i.e., recommending future experiences and training) also depends on learner assessments. Content brokering will support just-in-time learning and sequencing of learning events. Competency models will enable content to be tailored to the individual learner's needs.

Training applications like GIFT operating in the TLA environment will both consume learner data available through TLA and provide learner data as they complete training. It will be challenging to insure that all training applications will be able to both obtain necessary learner data from TLA as well as insure that they all output learner measures that can be used by other applications within TLA. This chapter explores some of the challenges of integrating a training application like GIFT into the TLA. This includes discussion of the discovery and development of methods to assess competency based on xAPI statements and recommendations for augmenting xAPI statements to facilitate interoperability among training applications and the TLA through methods such as semantic analysis.

The Evolution of the Current Defense Training Architecture

The rapid growth of the World Wide Web in the early 90's opened new opportunities for delivering computer-based instruction (CBI). As this medium was increasingly used in both higher education and business, learning management systems (LMSs) were created to facilitate the delivery of CBI (Kamel, 2008). By the year 2000, there were over 100 LMSs on the market though the vast majority of the market was dominated by a handful of systems (Falvo & Johnson, 2007).

There is no industry standard for what should or should not be in a LMS; however, there are some core functions that most of them possess. These include the ability to schedule courses and enroll students, support collaborative learning, store and deliver content, support learner evaluation/certifications, manage student records, and support career planning (Hunke & Johnson, 2006; Kamel, 2008).

In 1999, the Army published its requirements for the Army Learning Management System (ALMS) and began production in 2004. By the mid to late 2000's, all services and the Department of Defense (DOD) were looking to adopt LMSs for the delivery and management of online and blended courses (Hunke & Johnson, 2006; Kamel, 2008; Shanley, et al., 2012; Graul, 2012).

Though LMSs are widely used by government, academia, and industry today, they have their limitations. LMSs were developed to support enclosed learning systems on which the business of education depends. Specifically, LMSs control access to, schedule, and deliver curricula and by extension degrees and certifications earned.

Increasingly, it is recognized that training and education take place continuously in peer-to-peer interactions, during the execution of one's job, and in any number of self-development activities. Through the internet, individuals can access how-to videos, blogs, forums, webcasts, etc., to get answers to questions or training on just about any topic. The current generation of LMSs are not equipped to monitor or manage any of these kinds of learning interactions. There is a need for LMSs to evolve.

A learning architecture helps to define the basic functionality of the next-generation integrated learning environment (Kamel, 2008). The DOD Advanced Distributed Laboratory (ADL) has described just such an architecture. It is known as the Training and Learning Architecture (TLA).

TLA Overview

The TLA is a vision of the next-generation integrated learning environment. As mentioned previously, the capabilities of current LMSs define and constrain e-learning and blended learning environments today. LMSs provide standardization and structure to the way learners interact with content (Kamel, 2008). Unfortunately, these attributes work against a learning environment that seeks to deliver training in a way that uniquely adapts to each individual, an environment envisioned in the TLA.

The TLA includes five basic functions: experience tracking, learner profiles, content brokering, and competency networks (Johnson, 2013). To facilitate innovation and make it easier and less costly to maintain and improve, the TLA will be based on non-proprietary, open-source approach to delivering services.

The first function of the TLA, experience tracking, is all about being able to monitor learner activities wherever they may occur. In most LMSs today, measures of learners consist of summary scores, course completion, hours of instruction, and the like. These are essentially the same measures that have been recorded since formal education began. With the massive growth of online content, learners often spend a considerable amount of time outside of the LMS before, during, and after a course to prepare for the course, improve their understanding, refresh their training, seek tutoring or support on challenging topics, and even to share their own knowledge with others. By monitoring these kinds of behaviors, quite a few interesting outcomes are suddenly possible. For example, it would be possible to determine what course content is particularly challenging for students. It would also be possible to then automatically tailor training for those individuals. One could also determine what material is retained well and what seems to be most easily forgotten.

The ADL has created a specification for experience tracking called the xAPI, which is a non-proprietary specification for tracking and storing experiences across learning platforms (e.g., simulators, virtual worlds, web content, mobile devices, games, and observer-based measures). The learner's activity stream (a series of xAPI statements) is stored in a JavaScript Object Notation (JSON) database called a learning record store (LRS). Each statement includes the actor, verb, object, and optional information about results, context, etc. The power of xAPI is that it can record human performance at both the micro and macro levels. For example, a single keypress or behavior can be an xAPI statement, but it can also be a statement that the individual completed a course, received a certification, etc.

The second capability of the TLA is learner profiles. A learner profile is a map of the learner's background, experience, knowledge, and traits. It goes beyond a simple student record held in current LMSs, which primarily include courses completed, grade point average, etc. The learner profile in the TLA would include any information about the learner that may impact how and what training should be provided to the student. For example, cognitive abilities like intelligence, reading speed, and reading level. It also includes prior experience and knowledge in different domains like math or human psychology, small unit tactics, or how to operate a specific system. This profile is clearly dependent on the measurement of learners and so a LRS is a key enabler of the development of learner profiles. A learner profile is more than just a collection of measures. For example, over time, learners forget and lose skills when they don't have an opportunity to practice. A learner profile needs to take into account skill and knowledge retention over time if it is to be accurate.

The third capability of the TLA is competency networks. Each branch of the military has defined competencies that they need in their respective workforces. Some competencies may be common across all service members while others may be very specific to a single specialty. Competencies are defined by organizations as are the training and measures needed to develop and maintain those competencies. Competencies are complex and are usually developed over years through a combination of formal education, mentoring, job assignments, peer interactions, and more. A competency network is a way of representing those types of experiences and how they feed into the development of various competencies. By mapping the learner profile to the competency network, it will be possible to determine the competencies of the individual. The TLA will provide each service with a way to represent the training and measurement of their respective competencies.

The final TLA capability is content brokering. Content brokering has to do with tailoring the delivery of content to individual learners. Content brokering is dependent on knowing what the learner has done (via experience tracking), what the learner currently knows (via the learner profile), and what the learner needs to be provided with (via the competency network). Content brokering uses content registries and repositories to make training recommendations. Because so much content is now available outside of the walls of most LMSs, tools that enable machines to find and understand learning content automatically will play a big role in content brokering. These tools will do more than simply perform a semantic analysis, they will consider who uses them, how they are rated, and which ones seem to provide the biggest benefit to learners.

Current Learner Assessments in GIFT

Although GIFT was developed prior to the TLA, the assessment mechanisms within GIFT are compatible with various aspects of the TLA. While the TLA may be concerned with experience tracking across a range of formal and informal learning activities, GIFT is also using evidence-based assessment techniques to determine the current and projected performance of each learner and team based on their historical performance in a domain and based on behavioral markers and physiological measures, which indicate cognitive, affective, and physical states of learners.

In many ways, the assessments and services proposed for the TLA mirror those in GIFT. The key difference being that where the TLA is concerned with the progression of an individual at the curricular and career levels, GIFT is concerned with the progression of a learner at the course or lesson level. To do this, GIFT uses its own versions of experience tracking, learner profiles, content brokering, and competency networks to know where the learner's knowledge and skill are at each point in the course (learner module), where the learner needs to go (domain module), and what methods and content (pedagogical and domain modules) to employ to help the learner get there (see table 1). Figure 1 illustrates how GIFT uses these assessments and processes to tutor an individual using what is known as the learner effect model (Sottilare, Brawner, Goldberg & Holden, 2012; Sottilare, Goldberg, Brawner & Holden, 2012; Sottilare, Brawner, Sinatra & Johnston, 2017).

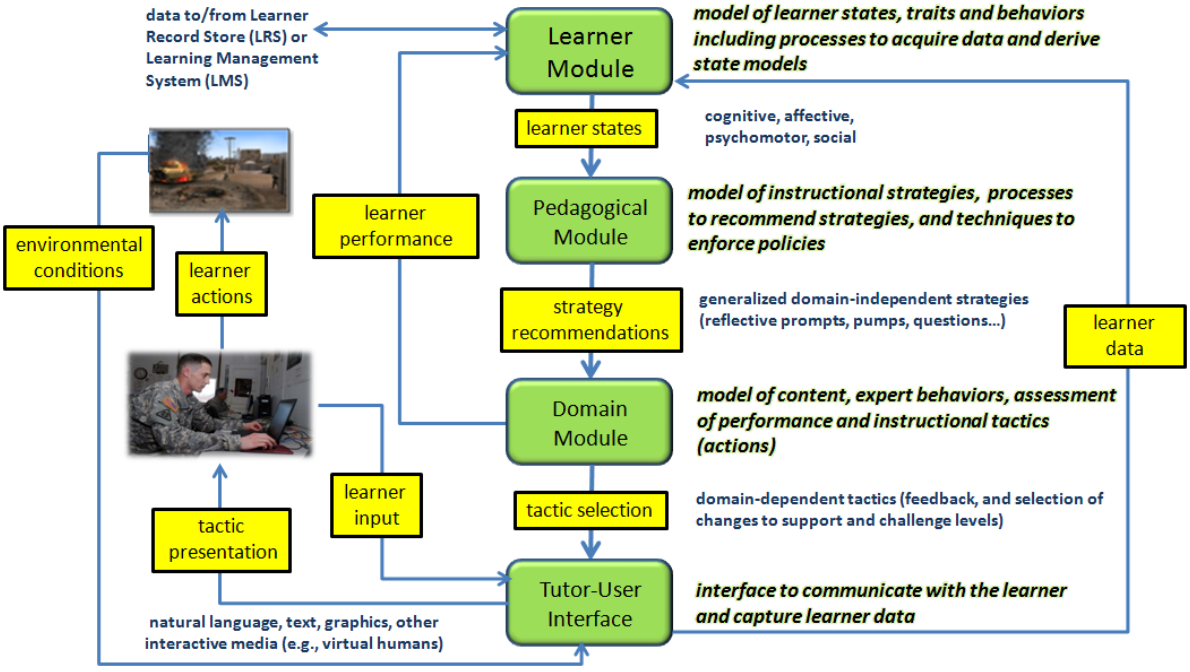


Figure 1. Instructional processes and architecture of GIFT.

GIFT is a framework that modularizes the common components of ITSs. These components include a learner module, an instructional or tutor module, a domain module, and a tutor-user interface. One of the main motivations for creating this framework was to lower the cost and labor needed to create ITSs by facilitating reuse of components and by simplifying the authoring process (Sottilare, Goldberg, Brawner & Holden, 2012).

The learner module represents the current state of the learner and also tries to predict future states; however, the learner module is not typically conducting assessments in GIFT. Assessments are done by the domain and sensor modules which pass their assessments to the learner module. The learner module uses those assessments as well as demographic and historical data about the learner to provide a classification of the learner's cognitive, affective, psychomotor, and competency states.

Right out of the box, GIFT can collect student measures in several ways. First, GIFT uses surveys for soliciting responses from learners. Survey questions can be used to collect demographic information or administer standard psychometric questionnaires like the NASA Task Load Index. Survey questions can also be used to test comprehension or knowledge either pre, post, or during training. These assessments of

comprehension are done in the domain module and then passed to the learner module for management/maintenance of learner state representation (Sottilare, Brawner, Goldberg & Holden, 2012; Sottilare, Goldberg, Brawner & Holden, 2012).

GIFT can also log all student keyboard and mouse actions that occur when learners interact directly with the tutor-user interface. When GIFT passes the learner to another training application, like a simulator, GIFT can collect learner interactions with those applications via an application programming interface (API). In fact, GIFT comes with ready-made gateway modules for some popular applications like Microsoft PowerPoint and VBS2. Once again, when these interactions are used to assess learner's understanding of concepts, knowledge, or skill by the domain module, those assessments are passed on to the learner module to update and maintain a representation of the learner state.

Finally, GIFT provides a standardized means for collecting data from a variety of commercial sensors that record physiological (e.g., electroencephalogram, electrocardiogram, electromyogram) and behavioral (e.g., eye-tracking, Xbox Kinect, force transducers, accelerometers) data. Sensor data a filtered, segmented, and/or extracted by GIFT's sensor module to provide a basic assessment of the raw sensor data. This basic assessment can then be used by the domain module to assess the learner's state, and this state assessment is then passed to the learner module. Raw sensor data are also logged for post-training analysis.

A simple framework for understanding how different types of assessments are used by GIFT is depicted in Figure 2. This framework divides assessments across two dimensions. First, assessments are divided into pre-training and in-training categories and second, they are divided into domain dependent or independent categories (Goodwin, et al., 2015).

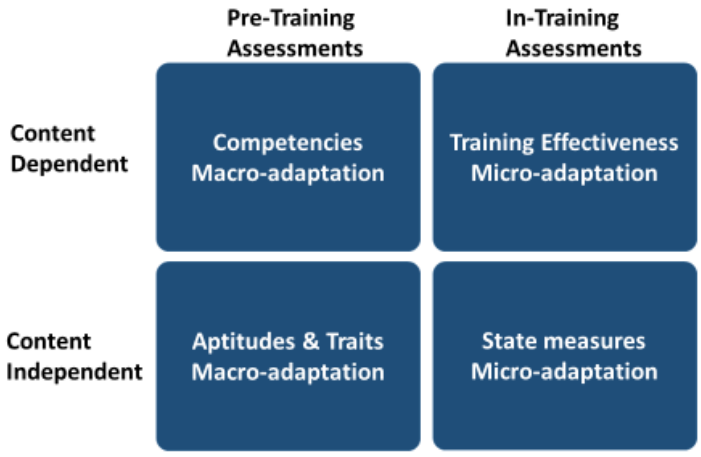


Figure 2. Conceptual framework for assessments.

As can be seen in this figure, assessments that are completed prior to training primarily influence larger macroadaptive strategies that GIFT would employ. For example, if training on the operation of a tactical radio was being delivered to a medic, examples and exercises would be relevant to the medic mission as opposed to the infantry mission and so forth. Assessments done in training would primarily influence micro-adaptive tactics such as whether to increase the difficulty of the instruction, whether to provide hints or feedback, etc.

Because GIFT and the TLA both engage in learner assessments, content brokering, learner modeling, and experience tracking, it is necessary to determine how these two systems would work together in a coordinated fashion. These topics are discussed in the following sections.

Integrating GIFT with the TLA: Experience Tracking

The xAPI enables a revolutionary new capability for *experience tracking* that forms the core of data sharing and interoperation in the TLA. Experience tracking refers to encoding and storing in a sharable manner some fine-grained information about actions a learner takes or events that impact a learner during a simulation, across a college course, or simply while consuming a video or text. In turn, understanding what individual learners experience from moment to moment can provide the grist to choose content that relates to the experienced context, promptly identify and respond to key moments, or proactively create conditions for learning. As such, the experience tracking approach colors the TLA perspective on competency management, learner modeling, and content brokering.

General challenges associated with experience tracking include scope, tractability, and privacy. Scope refers to the questions surrounding what types of data should be included in experience tracking, what level or granularity, what frequency of update, and so on. While recording more data certainly creates more raw material to work with, indiscriminate capture increases the difficulties related to tractability and privacy. Tractability refers to issues introduced with storage size, computation speed, network latency, human attention if required, and any other resource that can be strained by large amounts of data. Experience tracking can be made more tractable with selectivity about what experiences are useful to store or using batching methods. Privacy concerns relate to the storage of experiences that reasonably impact learning but might not be desirable to share publicly, such as arrival at a real-world location or failure on a high-stakes test. Methods being explored to protect privacy include identifying opportunities to store categories of information rather than unneeded specifics and introducing access control that limits who can read recorded experiences.

Specific to adaptive training systems, there are exciting instructional design challenges surrounding how adaptation that responds to a different training system can improve learning. What is acceptable or helpful in terms of adapting learning to respond to a past experience in another system, or even shifting control over the learning experience between systems? At the technical level, an important challenge is the need for one adaptive system to understand the experiences recorded by another training system. Maximally effective adaptation requires knowing the learner's past experiences, which are stored in recorded xAPI messages. However, it can be easy to share this valuable knowledge about experiences but hard to understand what someone else has shared. Some reasons for this are varying interpretation of the standard xAPI vocabulary and the ease of creating new vocabulary to encompass every situation. The semantics of a message might also vary subtly with context, such as the sequence of past experiences that led up to it or even which version of the software wrote it. As xAPI finds widespread acceptance, how to use each part of the vocabulary is being defined and agreed upon in communities of practice. However, new users and new use cases will continue to introduce variations.

GIFT is capable of writing and reading xAPI messages. Examples include messages for GIFT course completion as well as for responses to tests or surveys. The specific xAPI statements generated by GIFT are currently determined by course authors. Because there are very few consumers of xAPI data streams, course authors have little guidance as to what kinds of experience statements should be produced. For the TLA to make use of the experience reporting capability of GIFT or other adaptive training systems, it will be necessary for the TLA to be able to make requests to those systems. Additionally, it will be helpful for the ADL to provide some guidance on what kinds of experience streams it needs. For example, ADL is working to determine the balance between recording domain-specific data versus measures that would have more

general, domain-independent uses. Another challenge is developing standardized ways of reporting measures within specific training domains. There are always a variety of ways to report skill levels, knowledge, and experiences relevant to specific domains. To make it easier to compare and aggregate such measures across delivery platforms, standards are needed. Currently, the ADL is using communities of practice to develop standards for reporting activities, context, and processes so that a common vocabulary can be used by all practitioners.

Integrating GIFT with the TLA: Competency Management

The TLA uses common information about competencies that can help to coordinate teaching and training across systems. Competencies in the TLA refer broadly to skills, knowledge, abilities, and other targets of learning. TLA competency information includes human-readable descriptions, machine-readable definitions, and relationships between competencies such as levels and prerequisites. Clearly defining competencies helps make learner assessments portable across training systems.

Typical challenges surrounding competency management include agreeing on the meaning of a competency and who is allowed to define the meaning. Many different interpretations and weightings of component factors are valid for different uses. For example, a competency reflecting knowledge of a cyber intrusion tactic for the general population might require knowing that it exists or knowing rules to avoid the attack. For more specialized learners, related but different competencies might need to be defined that require understanding of underlying principles or even ability to carry out the tactic. In addition, competencies may be associated with different learner performance across populations, which complicates leveling and norming. As such, competency definitions may be created that are so broad they become vacuous or so specific they are not amenable to reuse across the TLA. Human-readable and machine-readable competency definitions must be carefully aligned. Also, careful judgment may be needed in maintaining the competencies because any change to a definition might disrupt another system that uses the same competency, while excessive versioning or division of the competencies might break links between systems that should survive that change.

Within adaptive training systems that use managed competencies, different structures may be needed to make the competencies applicable to different models of learning. Systems may require different relationships to be defined or interpret relationships differently (e.g., creating correlations between skills as historical data or for the purpose of machine learning and inference). They may need specialized data about relating competencies to specific content (e.g., which skills contribute to good performance in a simulator) or about content characteristics (e.g., what modes of presentation are appropriate). So, a key challenge is selecting the data to store for each competency and then collecting that data. The large number of competencies and all their details are unlikely to encoded completely or correctly, or they may become out of date in a process called *concept drift*. Because of these challenges, it is likely that adaptive training systems require a way to impute or learn and refine competencies.

In considering the prospect of GIFT making use of competencies that are defined and managed in a TLA-enabled location outside of GIFT, policy and technical challenges arise. Are GIFT competencies defined by a single central body with authority or are unit-specific differences expected? Are competencies defined by one military group allowed to be reused, reinterpreted, or even changed by outsiders? Finally, it is known that the existence and definitions of some competencies may not be published or shared, for example, when they are classified. The TLA supports such an environment by allowing multiple components to manage competencies separately, so that for example an unclassified component might manage most competencies while a component on a secured system could manage classified competencies. However, such a separation will introduce technical challenge surrounding replication and coordination of data for processes that need secure access.

Integrating GIFT with the TLA: Learner Profile

Learner models contain data about assessed proficiency or mastery of competencies as well as traits, demographic data, preferences, learning goals, and transient states of the learner.

Broad challenges of learner models include those that are well discussed in the literature as well as interesting new challenges introduced by the TLA focus on fine-grained experience tracking. First, there is a need to create unified measures to express each learner state or trait, and translate between measures. Since it is unlikely that every different training system will use the same scales or metrics to describe learner characteristics, methods must be created to let different measure interact. Does a four out of five on one scale equal an eight out of ten on another scale? Where does "meets expectations" fall on either scale? Standardization in the TLA aims to let systems answer some such questions automatically. Next, making sense of learner experience records requires rolling up raw data from one or many experiences into actionable information. This requires understanding different models of change, including differing stages to express how learning takes place and differing models of skill or knowledge decay. Finally, variations in processing across training systems should reasonably lead to varying levels of trust in learner model contents that are shared from other training systems. It should be possible to evaluate information in a learner model based on its recency, authority, and so on. Therefore, systems participating in the TLA require architectural capabilities to tie each learner competency estimate back to the components and the evidence that produced it.

An important challenge specific to adaptive training systems is the *cold start* problem, where computer systems must take time to identify learner traits necessary for adaptation. The TLA is specifically designed to mitigate cold start difficulties by sharing the needed information with participating systems. A second challenge is the often negative perception of assessment associated with high-stakes testing. However, the value of the TLA does not require high-stakes tests or formal assessments. The TLA may support more formative assessment that acts to address some concerns about time spent on testing. Finally, there exist a challenge surrounding learner models that may contain personally identifying information or otherwise sensitive information such as records of failures on important training events. As described previously, personal control over private information or intelligent filtering such as sharing the lack of a success to date, rather than a definitive failure, could provide approaches to address such challenges without reducing the important value of a shared learner model.

In the GIFT framework, modeling learner characteristics outside of the established GIFT model suggests a possible opportunity for automated intake of new learner model definitions. However, the learner characteristics in GIFT are by design very general to all training domains. As a result, it may be that design changes will be needed before GIFT can incorporate new shared characteristics that let GIFT take advantage of the learning context other systems know. For example, it may be desirable to introduce a layer of interpretation that can translate learner model characteristics for GIFT instead of, or as an intermediate step toward, adding the characteristics as first-class members in the GIFT learner module.

Integrating GIFT with the TLA: Content Brokering

Content brokering refers to recommending future experiences and training based on learner goals, characteristics, and assessments.

Similar to competency management, managing characteristics of content is difficult to do at scale. Metadata and paradata describing each piece of content needs to be authored and stored. As we have argued, such automated collection and maintenance of such data is likely to play an important role in TLA-enabled train-

ing systems. Automation will help ensure that data is accurate and up to date, which is of increased importance when different systems need to coordinate to understand and recommend or broker learner experiences.

An interesting research question surrounds the assembly of a unified learning experience out of the atoms provided by different adaptive systems. Content brokering is likely to take advantage of, and be challenged by, new modes of learning such as second-screening or switching between systems when an experience is incomplete, as opposed to linear completion of a single recommended learning pathway. Unified language and surface presentation of instructional content is likely to require precise content descriptions that are not currently available. Some even argue that unified user interfaces, iconography, or fonts might be needed to avoid distraction and extraneous cognitive load caused by switching between training systems.

The granularity of content to be brokered is an interesting question. While humans are able to make "close enough" fits between learner needs and content, or quickly identify atoms of content to a very fine-grained level, the same can be difficult for automated systems. For example a human may find it appropriate to direct a learner, "read the first two pages, the last three are not related to our work now." Adaptive training systems need a way to link deep into content and markup the different competencies related to fractional portions of content that is experienced. This is an interesting challenge because content portions that are viewed or need to be viewed may be continuous and have graduated effects rather than discrete. As an example, watching the first 60 seconds of an instructional video may have different effects from watching the first 90 seconds. The video may or may not be possible to break into discrete segments that have beginning and end points a machine can identify. In a training simulation, it may be desirable for a learner to experience one particular path or a few out of thousands of possible paths. Content brokering needs a method to understand these impacts and select, suggest, or influence certain paths at a fine grain that expert human instructors can achieve.

GIFT may provide a central method for the TLA to start and control content in adaptive systems. However, this GIFT approach to content brokering currently requires a GIFT-specific interoperability program that costs developer time to create. The GIFT framework also assumes that content it brokers within its framework will give GIFT certain types of control over the learning experience, such as allowing GIFT to terminate running content. This kind of centralized control makes sense in GIFT and is used in useful ways such as switching to another activity without leaving too many windows open. However, not every system that participates in the TLA will find it possible to use this structure. A minimal set of direct interoperability controls should be identified and possibly modularized into a reusable content brokering component that could reduce the developer effort to participate in GIFT content brokering.

Conclusions and Recommendations

The framework being proposed here for the integration of GIFT (or any adaptive training system) and the TLA sees the TLA as handling management of training and education at the course and above levels while a system like GIFT handles management of training at the course and below. For example, the TLA might determine that a learner needs some block of training in GIFT to develop or maintain a particular competency and so it would in essence bring the learner to the GIFT classroom for training. As part of this handoff, the TLA would provide GIFT with a record of relevant prior experiences and training of this learner. At the conclusion of training, GIFT would return the learner to the TLA and would update the learner profile accordingly.

As discussed previously, there are some challenges that need to be addressed in order for this framework to become reality as shown in Figure 3. First, for the TLA to deliver the learner to the GIFT classroom, the

TLA needs to be able to know about the training available in GIFT and how it maps into various competencies. To the degree that there are not clearly defined methods for identifying competencies and defining competency networks, it is not clear what the link between a GIFT course and a competency network would be. Even assuming such competency networks exist, there is still the challenge of determining how to evaluate the content of a GIFT course so that it could be mapped to specific nodes on the network.

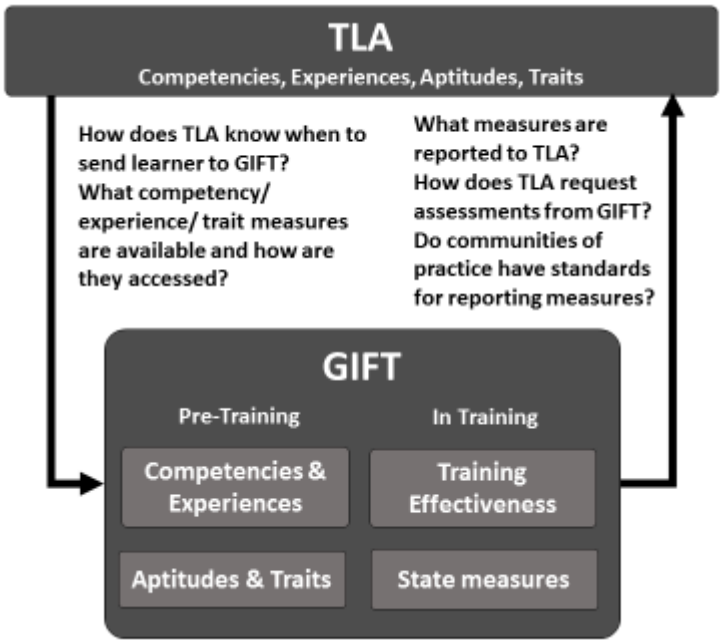


Figure 3. Challenges for integrating GIFT and TLA.

In terms of experience tracking, GIFT can generate xAPI statements that can be fed into an LRS. So, on the surface, it would appear that GIFT is already largely TLA-compliant. The biggest challenges that exist with regard to experience tracking are not specific to GIFT, but are general challenges of experience tracking. Specifically, standards need to be developed for reporting of activities within specific training domains to facilitate comparisons of measures across applications. Other issues involve determining the granularity of learner activities that should be reported and the identification of critical contextual information that is needed for proper interpretation of activities. As communities of interest develop conventions and standards for experience tracking, GIFT will need to comply with those standards, but that shouldn't be a difficult to do.

In conclusion, it is clear that GIFT and other adaptive training systems must be part of an integrated learning environment like that envisioned in the TLA to realize their full potential. In this environment, GIFT will know about learners as soon as they enter the GIFT classroom, maximizing GIFT's training efficiency by eliminating the need for it to spend time interrogating the learner before delivering its training. This environment will also enable GIFT to adapt training across training venues or modalities. For example, suppose a course includes a sequence of classroom training followed by simulation training followed by live training. If learner's activity streams are being recorded in the classroom, GIFT would be able to adapt the simulator session based on their classroom performance. Similarly, if performance is tracked during live training, GIFT could both use that data to evaluate the effectiveness of the training it delivered in the simulator and even to recommend additional remedial simulator training when needed.

References

Falvo, D.A. & Johnson, B.F., (2007). The use of learning management systems in the United States. *TechTrends, 51*, 40–45.

Goodwin, G., Johnston, J, Sottilare, R., Brawner, K., Sinatra, A., Grasser, A. (2015). Individual Learner and Team Modeling for Adaptive Training and Education in Support of the US Army Learning Model: Research Outline (Special Report 0336). US Army Research Laboratory: Aberdeen Proving Ground, MD.

Graul, M. (2012). Framework for adaptive learning Content management and delivery (FALCON – Report AFRL-RH-WP-TR-2012-0192). Air Force Research Laboratory, Wright Patterson Air Force Base: Dayton, OH.

Hunke W. & Johnson, R. (2006). Learning Management Systems White Paper (Report TAT-09429). Information Assurance Technology Analysis Center: Herndon, VA.

Johnson, A. (2013). The Training and Learning Architecture: Meeting the Needs of the Next Generation of SCORM. Slides from a Webinar. Downloaded from: http://adlnet.gov/wp-content/uploads/2013/02/TLAWebinarFeb2013-1.pdf on 4 Jan, 2017.

Kamel, M.N. (2008). Learning Management Systems: Practical Considerations for the Selection and Implementation of an E-learning Platform for the Navy (Report NPS-GSBPP-08-011. Naval Postgraduate School, Graduate School of Business and Public Policy. Monterey, CA.

Shanley, M.G., Crowley, J.C., Lewis, M.W., Straus, S.G., Leuschner, K.J. & Coombs, J. (2012). Making improvements to the Army distributed learning program (Report MG1016). Arroyo Center, Santa Monica, CA: Rand Corporation.

Sottilare, R., Brawner, K., Goldberg, B. & Holden, H. (2012). The Generalized Intelligent Framework for Tutoring (GIFT). US Army Research Laboratory.

Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017. DOI: 10.13140/RG.2.2.12941.54244.

Sottilare, R., Goldberg, B., Brawner, K. & Holden, H. (2012). A modular framework to support the authoring and assessment of adaptive computer-based tutoring systems (CBTS). In Proceedings of the *Interservice/Industry Training Simulation & Education Conference*, Orlando, FL, December 2012.

CHAPTER 4 – Enabling Intelligent Tutoring System Tracking with the Experience Application Programming Interface (xAPI)

Andy Johnson[1], Benjamin D Nye[2], Diego Zapata-Rivera[3], and Xiangen Hu[4]
Advanced Distributed Learning Initiative[1], University of Southern California,
Institute of Creative Technologies[2], Educational Testing Systems[3], University of Memphis[4]

Introduction

Two trending areas in learning technology are intelligent tutoring systems (ITSs) that can create truly adaptive instruction for individual learners and teams and increased data interoperability through use of the experience application programming interface (xAPI). xAPI, a set of application programming interfaces (APIs), allows a common data structure that applications can use to share information. Developed by the Advanced Distributed Learning (ADL) Initiative as a free, open-source, and highly extensible specification, xAPI offers a great number of benefits to an intelligent tutoring capability (ADL, 2016).

This chapter explores the possibility of implementing xAPI within an ITS or across ITSs. First, the current state of ITSs is reviewed to determine how domain-specific an ITS is and to see how difficult the barriers to implementation are to overcome. If ITSs can perform in an interoperable way, they can focus on a particular domain and optimize performance within that domain. At the same time, data should be accessible to enable those who could do powerful things with them. Next, data analytics are explored both in terms of what the field is doing as well as how ITSs are entering that field. This introduction into interoperability and data exposure provides the groundwork to look at the benefits and breadth of xAPI, with an eventual goal of semantic, not just syntactic, interoperability. Finally, the possibilities of ITS + xAPI has great potential. As part of this discussion, we explore solutions with an xAPI community of practice (CoP) and examine methods to expand xAPI vocabulary to capture achievements within ITSs and support assessment.

Related Research

When determining the viability of an ITS, one needs to realize that an ITS is often very tied to a specific task domain. Obviously, subject domains such as English can react to anything thrown at it or simply find relevant content on the Internet based on an individual learner's needs. Previously authored content may already exist. Learning content on its own cannot be a domain, it is far too dynamic. Activities in the "Learning" domain can include everything from essays to simulations to graphing to collaborative design. This gets to the central point made by Nye, Goldberg, and Hu (2015): learning tasks for a domain define how content is authored in that domain. This argument doesn't even get into other domains that are ripe for ITS improvement such as assessment (interactions or checks to determine understanding), measures (a criterion to determine whether a learner's state is at, below, or above expectations), motivation (internal and external factors that affect achievement), cognition (acquisition, processing, and using knowledge), affective states (attitudes, values, and emotions), and psychomotor (manual or physical skill acquisition).

The argument of learning tasks defining authoring practices has many facets to it. First, that a domain or subject area will have a subset of activities of the aforementioned "Learning" domain. This alone adds complexity of the notion of a generalizable ITS. Before proceeding to a further argument that ITS specialization is inevitable, we must define a learning task. A learning task from this authoring standpoint is defined as 1) having a distinct pedagogical state, 2) having dynamics during that task that are mainly or wholly derived from task state, and 3) including actions performed by the learner (or learners) (Nye, Goldberg &

Hu, 2015). As one could guess, there are many complexities in an ITS, which relies on both a pedagogical and learner state model. The level of effort to author also increases with feedback. The larger the domain of the ITS, the more complex.

The vision of a single, large ITS also loses viability when considering barriers to ITS adoption. Nye (2014) cites a number of these barriers. First, due to an explosion of data, it is extremely difficult to test which features of ITS are valuable by turning them on or off. Second, students still need to have access to and be motivated by an ITS. Access to technology cannot always be assumed, especially as the infrastructure needed for the ITS increases. Social expectations of an ITS also weigh heavily into consideration. In a similar way, the third barrier is proving value to teachers and administrators. Teachers need to be "in the loop" and reports of the student model need to be accessible. Finally, and related back to the authoring argument, usage of ITS is dependent on premade curricula. An ITS isn't a magic box that creates learning content, it makes existing learning content better.

All of these complexities make the possibility of having an ITS that contains all of a learner's needs regardless of domain, virtually impossible. A single large ITS isn't the solution for an ecosystem that can support all of a learner's need. Instead, distributed ITSs that can communicate with each other seamlessly are needed.

The Need for an ITS Framework for Interoperability

One only needs to look at the technology landscape to realize the movement to distributed services. Smartphones and tablets with applications from multiple developers have replaced software bundles. Web surfers utilize a variety of services with a mentality of "if they don't work nicely with each other, I'm not going to use it." There is no doubt that ITSs will need to move in the same direction. Thus, it is no surprise that researchers in the area of ITSs have explored various methods for sharing student model information collected by various ITSs in different contexts with other systems. These approaches include student modeling servers or shells (Fink & Kobsa, 2000), distributed agent-based platforms (Vassileva, McCalla & Greer, 2003), distributed agents and servers (Brusilovsky, Sosnovsky & Shcherbinina, 2005; Zapata-Rivera & Greer, 2004), peer-to-peer systems (Bretzke & Vassileva, 2003), service-oriented systems (Kabassi & Virvou, 2003; Winter, Brooks & Greer, 2005). Although this work has resulted in platforms that can support the integration and sharing of student model information, scaling-up and human support issues may have prevented the widespread use of these platforms among groups outside the institution where these approaches were originally developed.

The biggest hurdle to these technology challenges is, of course, inertia. Many educational institutions have to deal with the challenge of consolidating student data that has been stored in silos across the organization (Daniel & Butson, 2013). Government is even more of a challenge with many more acquisition rules and layers of bureaucracy. Many software systems are large and were expensive to build, and are also extremely costly to maintain, which makes trying something new difficult. Despite these challenges, the field of ITSs would be best served to move to an open framework in the near term, rather than dealing with the predictable problems of increasing maintenance costs and old technology deprecation it would face if designed without interoperability in mind.

The goals of this interoperability should be not only syntactical compatibility between two systems that want to exchange data, but actual semantic interoperability. That is, the machines not only agree on what data looks like, but also agree in what they mean in a larger learning ecosystem.

Discussion

The Importance of Data Analytics

An area of increasing importance as technology advances is data analytics. Data can move faster through increasingly fast infrastructure with servers that have more capacity and are able to process more information than ever. We see this in ITSs as learning analytics and educational data mining methods are frequently used to extract information from large datasets to inform the creation of adaptive educational systems and provide teachers with actionable data to support student learning (Baker & Inventado, 2014; Arnold, 2010). There are many benefits to performing this type of analysis. Learning analytics can help with analyzing student data and creating models that can be used to identify students at risk and suggest possible interventions (Daniel, 2015). While identifying at-risk learners is important, so too is identifying those overachieving. Such learners are ready for more challenging tasks or potentially leadership activities. Models don't have to focus on simply performance, but skill sets as well. Matching a person's skills and abilities to job functions is a historically difficult problem in the DOD (and likely elsewhere). Using data to infer better matching would be a boon to employers. Data analytics enable some interesting social engineering possibilities. By displaying the results of analytics, dashboards can be created and exposed to learners as well, thus increasing both competitive and collaborative possibilities in a motivational/gamified environment.

The field of ITSs can benefit from some lessons learned in distributed learning content, especially with regard to exposure of data for analytics purposes. Around 2010, the ADL Initiative launched a series of surveys, interviews, and questionnaires surrounding the Sharable Content Object Reference Model (SCORM) to prepare to create a new specification that could fill gaps as reported by current users and tool providers. While SCORM was very successful in bringing a standard to distributed learning, the aspects of "big data" and even access to data were raising needs that SCORM did not meet. Respondents reported a need for students, teachers, administrators, and authors to have access to data. They wanted ways to expose the data directly, and not be reliant on a proprietary user interface (UI). In addition, the scope of the data needed to change. The interest in data across a learner's lifetime, how they perform in a group, and the varying types of assessment they take were all "need to haves" in this next specification (ADL, 2011). While SCORM was successful at aligning disparate learning management systems (LMSs) and their data, the requirements have aged 20 years, the field of ITSs would be better off looking toward lessons learned and the desired direction of those deploying distributed learning solutions.

xAPI, Profiles, and Vocabularies

The breadth and depth of what xAPI can track is outstanding. It can track different granularities of learning events by selecting an option within a multiple-choice question to graduation from a University program. xAPI's range is also diverse in that it can track motions done by a solider during training using wearable technology or it can be used to record a conversation that happened between professors and their students. The reason is that xAPI has abstracted out many of the key fields in recording data. This flexibility has allowed applications to track any type of data that one can describe in language or in an attached file. Whether it is interacting with a digital publication; tracking the pauses, starts, and skips in a video; gathering training data from a simulation; or grabbing physiological data from wearable technology, xAPI can record events about learners in nearly any application. At the same time, xAPI provides details on *how* to represent users and other information, making it very flexible to build new ontologies. In addition, as xAPI is human readable (the names and values of the data format are intentionally based on natural language), barriers to entry are lower and creation of best practices is easier. A final benefit is that applications can and have been

developed that allow offline interactions, such as a technical evaluation of performance in an operation, to be integrated into an application post hoc.

xAPI provides a structure any system or application can get "back to". As previously discussed, "stove-pipes" of data are a burden to doing high-value analysis on performance. In most cases, even data migration is not possible because of the nature of the data being only accessible through a specific interface. xAPI reduces the barriers causing these silos in two ways. First, all xAPI data are exposed through querying controls built directly into the specification. To be xAPI-conformant, a Learning Record Store (the name of the data storage of xAPI) must allow access to all data (provided that user account has permission to see that data). While this may sound trivial, the inability for most systems to produce output that is usable by any other application is widespread in government, academia, and beyond. Second, xAPI uses a simple data format called JavaScript Object Notation (JSON) used widely in industry that can go anywhere – mobile, simulation, Internet of Things (IOT) devices, sensors, etc. By using a general transport format rather than a specific database schema, any application's database can easily export to xAPI or import xAPI. This alleviates the need for databases to "talk" to each other, which is important because the logistics around security and person-hours necessary for every single integration can quickly make collaboration unscalable.

The collaborative nature of xAPI means it is flexible as situations change and industry best practices change. Because xAPI doesn't incorporate its own vocabularies, CoPs can form around common interests and create their own best practices independent of the specification itself. It is expected that many CoPs exist, in particular to cover the variety of disciplines that want to track data. In addition, profiles, collections of vocabularies that the CoPs are likely produce, can be created and validated against in the same way conformance testing can be done on technical specifications. ADL is in the process of documenting processes and procedures for CoPs, profiles (and conformance testing of them), and vocabularies. The openness of the spec and CoPs also allow for the diverse requirements necessary in assessment. Organizations, states, and especially countries have very different requirements in regard to privacy, security, data tracking, and record keeping. xAPI offers support by being flexible in structure and in its use of identifiers rather than English tokens.

Recommendations

It is with this collaborative spirit that an ITS CoP needs to be developed and incorporated into the Generalized Intelligent Framework for Tutoring (GIFT). Not as an ironclad set of rules to be followed when tracking performance or assessing with the ITS community, but rather a common understanding of vocabulary for all possible implementation points and allowing organizations to create their own distinct solutions. While solutions may be distinct, developing a profile for ITSs would allow the common description of ITSs to have a clear path of how one ITS can interface with another. This interface is the key to assessment interoperability in ITSs. Research has already been done around evidence-centered design (ECD; Mislevy, Steinberg & Almond, 2003). ECD offers a principled approach for designing assessments based on the principles of evidentiary reasoning. It can inform the development of a vocabulary for assessment in the field of ITSs. The ECD community has developed tools and a vocabulary to refer to assessment design components (ECD Wiki, 2016). It is hoped that by the time this book is published, leaders in ITSs and xAPI will have established a CoP to begin this collaboration, in particular in the area of assessment. ADL has worked alongside many xAPI CoPs as they have developed, most notably, the cmi5 profile. The cmi5 working group has created a profile related to the traditional LMS model and combines the richness of xAPI with the structure of a basic learning assessment model. It has already being developed in Learning Record Stores and authoring tools. There is no reason that ITSs couldn't follow a similar path to success.

The best place for this CoP to start is with the recent recommendations of the US Army Research Laboratory team who created GIFT. Sottilare, Long, and Goldberg (2017) highlight five information classes that focus

on domain competence. The more accurate the competency modeling, the more effective an ITS can tailor or adapt instruction. GIFT and xAPI alignment needs to occur in five main areas. First, fine-grained achievement data, such as answering specific questions, navigating through certain sections, or interacting with components would be valuable to track in GIFT to make smarter decisions regarding content flow. Second, calculating duration will provide better metrics when compared with learner outcomes. This allows questions like "is a single 5-hour session more effective than 5-1 hour sessions" to be answered with data. Third, having a vocabulary and accompanying GIFT mechanism to allow a learner's to assess the quality of their experience would provide valuable feedback about the content itself. Fourth, an effective model for competencies and their decay is needed. This allows for notifications for refresher training as well as assessment of which strategies can allow for the least amount of decay. Finally, an assessment profile is needed within xAPI to track both formal and informal learning experiences. This profile goes beyond just a couple features and would have specific rules for usage. Implementing GIFT with boosted xAPI capabilities will provide ITSs the dynamic capability to meet the needs of 21st century learners.

References

Advanced Distributed Learning (2016) Retrieved November 1, 2016, from https://www.adlnet.gov/xapi/.

Baker, R. S. J. D. & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds), Learning analytics: from research to practice. Berlin, Germany: Springer.

Bretzke H. & Vassileva J. (2003). Motivating cooperation in peer-to-peer networks. In Brusilovsky, A. Corbett & F. de Rosis (Eds.), Lecture Notes in Artificial Intelligence: Vol 2701: User Modeling 2003 (pp. 218–227). Berlin: Springer-Verlag.

Brusilovsky, P., Sosnovsky, S., Shcherbinina, O. (2005) User Modeling in a Distributed Elearning Architecture. In: Ardissono, L., Brna, P., Mitroviü, A. (eds.) UM 2005. LNCS (LNAI), vol. 3538, pp. 387–391. Springer, Heidelberg.

Daniel, B., Big Data and analytics in higher education: Opportunities and challenges. British Journal of Educational Technology, 2015. 46(5): p. 904–920.

Daniel, B. K. & Butson, R. (2013). Technology enhanced analytics (TEA) in higher education, Proceedings of the International Conference on Educational Technologies, 29 November–1 December, 2013, Kuala Lumpur, Malaysia, pp. 89–96.

ECD Wiki. (2016) Retrieved November 1, 2016, from http://ecd.ralmond.net/ecdwiki/.

Fink, J. & Kobsa, A. (2000). A review and analysis of commercial user modeling servers for personalization on the World Wide Web. User Modeling and User-Adapted Interaction, 10, 209–249.

Kabassi, K. & Virvou, M. (2003). Using Web services for personalised Web-based learning. Educational Technology & Society, 6(3), 61-71. Retrieved October 10, 2006, from http://ifets.ieee.org/periodical/6_3/8.html.

Mislevy, R.J., Steinberg, L.S., Almond, R.G. (2003) On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives. 1, 3–62.

Nye, B. D., Goldberg, B. & Hu, X. (2015). Generalizing the Genres for ITS: Authoring Considerations for Representative Learning Tasks. *Design Recommendations for Intelligent Tutoring Systems: Volume 3: Authoring Tools and Expert Modeling Techniques* (pp. 47–64). US Army Research Laboratory.

Nye, B. D. (2014). Barriers to ITS Adoption: A Systematic Mapping Study. In *Intelligent Tutoring Systems (ITS) 2014.* (pp. 583-590). Springer International Publishing.

Sottilare, R.A., Long, R.A & Goldberg, B.S. (2017) Enhancing the Experience Application Program Interface (xAPI) to Improve Domain Competency Modeling for Adaptive Instruction. L@S 2017, April 20–21, 2017, Cambridge, MA.

Vassileva, J., McCalla, G. I. & Greer, J. E. (2003). Multi-agent multi-user modeling in I-Help. User Modeling and User-Adapted Interaction, 13(1/2), 179–210.

Winter, M., Brooks, C. & Greer, J. (2005, July). Towards best practices for semantic Web student modelling. Paper presented at the 12th international conference on artificial intelligence in education (AIED 2005), Amsterdam, The Netherlands.

Zapata-Rivera, D. & Greer, J. (2004) Inspectable Bayesian Student Modelling Servers in Multi-Agent Tutoring Systems. Int'l J. Human-Computer Studies, 61(4), 535–563.

CHAPTER 5 – Vision Statement: Navy Career Management and Training of the Future

Brent Olde
US Office of Naval Research

Introduction

This chapter presents a vision of future Naval career management and training[2] in which repeated performance assessments are critical to achieving the vision.

As technology become increasingly integrated into military systems, the ability to automatically gather operator performance data during day-to-day operations and training events has become easier and more cost effective. Collecting and analyzing these data provide an unprecedented capability to understand a sailor's duties and the impact of their training on their readiness to execute their mission. Having a performance assessment system that automatically identifies substandard performance and indicate areas for remediation would allow supervisors, instructors, and the individual to better understand their current capabilities and target interventions when and where needed. The data collected can provide training professionals with detailed objective metrics of their training effectiveness, allow comparative analysis of training efficiencies when training programs are modified, and provide military leaders with an understanding of their fleet's readiness. It is very rare to be able to directly compare the impact of training and education on job performance. For example, universities do not track all of their graduates and assess how well they do in their jobs after graduation, then feed that data back into their teaching to optimize their curriculum. The Department of Defense (DOD) is in a very unique position, as it takes individuals straight out of high school, trains them in their jobs, and could monitor their performance throughout their military career. At a superficial level it does this, but with the current technological trends, the ability to capture on the job performance metrics in a detailed manner opens up new and exciting opportunities to deeply understand a military job and customize the training to that job.

The Navy senior leadership have been promoting Sailor 2025 – Ready Relevant Learning (RRL), an initiative that will change the way they manage, train, and develop sailors (Burke, 2017; Faram, 2017). RRL relies on the adoption of the next generation of training technology, which typically requires the collection, storage, and mining of performance data. The Navy sees how industry is utilizing Big Data analytics to understand consumer behavior and grasps that similar techniques can be used to understand and improve how they manage and train their personnel. If done correctly, these data can be used to accomplish the following:

1. Assess and track individual/crew/unit/force performance and readiness.
2. Improving training effectiveness through the detailed analysis of training efficiencies.
3. Gather operational performance data to calibrate personnel selection and training.
4. Understand the relationship between various levels of training and their impact on mission execution.
5. Provide personalized, dynamic training requirements based on current operational needs.

The following vision statement provides an example of how a sailor's life could be in the future and alludes to the technologies needed to achieve that vision.

Paul is a couple years out of high school and is looking to get a better job. He has many skills but they were gained through self-study and experience but are not documented by a formal degree. He is interested in a job in the military, so he goes online to the Navy's recruiting webpage. He meets a computer avatar called the Personal Assistant for Life Long Learning (PAL3), which talks to him and gathers information about his formal education (his high school diploma), his past jobs, and personal interests, and administers a general assessment test[3]. The system then provides a list of jobs he is currently qualified for but also suggests he take an in-depth personality assessment[4] that would match his personality to the jobs that best suit him. He is a bit of an introvert, so although he qualifies for many positions, only a subset have a strong compatible satisfaction rating – so the system recommends he focus on those jobs. The system allows him to select current qualified positions but also other jobs he is interested in. The system has the capability to outline a career path to qualify[5] for any of those jobs (jobs he is not fully qualified for list the extra courses and training required).

Paul is interested in electronics and would like to work on a submarine. He also believes he already has some of the required skills, so he launches an in-depth electronics repair simulation[6] where he can troubleshoot a typical broken device. He is able to complete the task in the required time frame, so his profile is updated with his new verified knowledge, skills, and abilities (KSAs). He still does not fully meet the entry requirements for the job he is interested in, but the system indicates how he can gain the necessary knowledge and skills through several methods: college or vocational courses, apprenticeship/working in certain jobs, or some free online DOD classes. Over the next several months he takes a few of the free courses in math to qualify for the position. The system not only has the courses electronically[7], but has an intelligent tutoring system (ITS)[8] that provides assistance when he has questions or difficulty with the materials. The problems in the training are based on commonly experienced operational issues (mapped to the submarine electronics job he selected), so not only do the exercises proved realistic training but they provide an accurate preview of the type of work he can expect on the submarine[9]. The courses run on his personal mobile device and the device's camera uses facial recognition software, which provides feedback to the system on his emotional engagement (i.e., frustrated or focused) and also ensures he is doing the work himself and not getting too much help from outside sources.

Once Paul has completed all the required preliminary work, he goes to his local recruiting office. The recruiter accesses his profile and they discuss the opportunities available in the military. As a final step, he takes another assessment test (under controlled conditions) that verifies the pertinent KSAs needed for the job[10]. He passes this final test and is ready to enlist in the Navy; however, there are a couple months before he can report to the Recruit Training Command to start his basic military training (boot camp). During this time, PAL3 directs him to several available apps that provide material he will have to learn in boot camp (i.e., general orders of a sentry, recognizing Navy rank insignias, fitness standards, and general military training courses like sexual assault prevention and response). Paul studies hard and is well prepared for boot camp. He sails through graduation and quickly moves on to electronics school. Since he already has many of the skills needed, the system provides an accelerated curriculum, tailored to his specific training needs. The system provides a customized schedule, pacing the materials that need to be learned before a target graduation date. However, the schedule is only a recommendation. Paul can accelerate his training if he wants to put in more hours. He can test out of materials he has already mastered, or if he falls behind, it will adjust the schedule providing more time to complete. The flexible schedule tracks his progress and ensures he demonstrates full mastery of the topic areas before advancing. The system is also tied into a graduation tracking system, so any acceleration or delay in expected graduation date are tracked and reported.

The schoolhouse uses a flipped classroom model that allows the students to move through courses at their own pace. These courses contain basic computer content delivery but also virtual reality simulations that can walk students through immersive, job realistic problems to solve. There is an ITS available that provides nonjudgmental help anytime it's needed (Paul is a little intimidated by his instructors, so he really likes

going to the synthetic tutor). The school instructors can track Paul's progress with the learning management system (LMS)[11]; it automatically compares his progress to his peers and highlights his problem areas, thus allowing the instructors to spend their limited time providing in-depth, one-on-one assistance to those who request it or who the system flags as struggling. The system informs the instructor about specific material the student is struggling with and can even help set up an appointment for one-on-one instruction or remediation.

Paul completes his training early and is ready to join his ship; unfortunately, it is currently at sea and won't make port for several months. During this delay, his PAL3 continues to help him manage his career, it sends him questions of the day and assesses his responses, making sure the knowledge and skills he has developed do not decay and thus, he remains ready. With this downtime, the system focuses on further understanding Paul's career development goals and outlines the certifications, qualifications, and general path required to obtain those goals. With the time spent waiting, he is able to get a head start on some of his shipboard qualifications. Once onboard and actively working his new job, his daily completed maintenance actions are recorded. This provides the system with an objective assessment of the KSAs he is maintaining and the new ones he has acquired. Paul is driven and eager to advance, so he listens to PAL3's recommendations and takes some additional online courses.

This extra effort is tracked and can be seen by Paul's supervisors, who can access a supervisor's version of the system. The supervisor uses the information to track individual and unit readiness, conduct performance evaluations, and help recommend those individuals ready to take advanced training or promotion exams. The supervisor likes the performance evaluation feature. It provides an objective evaluation of a sailor's performance (i.e., tasks completed, difficulty levels of tasks, time efficiency, and errors) rated against their peers and only requires the supervisor to add their subjective input on how the individual has performed as part of the crew.

Since ships of the same class can vary greatly in the equipment they have onboard (variability abounds because the ships must rotate out of services to get updates), the tracked maintenance performance data in the systems is used to validate and update the current training content. The training material evolves as the fleet evolves; thus, as a new combat system become available, training for the older system is automatically replaced. Current sailors on the job and sailors going to that ship in the future are notified of the changes and are provided access to the new training. Since the system knows Paul is on a specific ship, he received customized training for that ship, with training that focuses on the common maintenance issues they have (based on historical norms and expected maintenance actions through predictive mean-time failure analytics).

The maintenance data are mined and the selection criteria for any fleet job is modified as the ship's duties evolve based on system changes, again maintaining accuracy to what is required as the position responsibilities evolve. Finally, the tracking system provides validation for civilian equivalency certifications that can be used if Paul decides to transition out of the military to a civilian career. This comes in handy when Paul, after been in the Navy for several years, decides to take a civilian job. He is able to use his civilian equivalency certification to gain a job. He settles down with his spouse and small child back in his home town. He does well in the civilian sector and learns new skills and professional certifications. He continues to access and update his Navy profile with his accomplishments with a notion of returning to service at some point in the future. The PAL3 system continues to provide helpful feedback and guidance so when he decides to return to the Navy, he is slotted into a career path that takes advantage of the new skills acquired in the civilian sector and thus maximizes his usefulness to the Navy.

Recommendations

The Generalized Intelligent Framework for Tutoring (GIFT) could be leveraged to achieve this vision; however, improvements need to be made in the following areas: providing users a method to gain insight into the performance, beyond the individual level (e.g., crew or unit), to provide readiness tracking; allowing the comparison of training effectiveness across similar tutoring systems; and providing guidance on the "right size" of competency development and how KSAs comprise the competency.

1. Dr. Brent Olde is a Commander in the US Navy. He is currently assigned as a Program Officer and Division Deputy at the Office of Naval Research, Human & Bio-Engineered Systems Division where he currently manages several training related science and technology programs. He received his undergraduate degree at the University of Missouri – Columbia and his PhD in experimental psychology at the University of Memphis, Tennessee, where his graduate work focused on the developed ITSs.

2. Although the vision is influenced by current Naval guidance, this chapter contains thoughts of the author and does not express any official Naval position.

3. General assessment tests could be the full or paired down version of the Armed Services Vocational Aptitude Battery (ASVAB) for enlisted or the Aviation Selection Test Battery (ASTB) for officer wishing to qualify for flight training.

4. In-depth personality assessments, like the Army's Tailored Adaptive Personality Assessment System (TAPAS) program. This type of assessment measure a persons' "fit" for a job. High school and college career placement centers have be using these tests for years to provide guidance on how much one would like a prospective career field. The Navy has a test like this to provide guidance on whether a sailor would be a good fit for life on a submarine.

5. This could be used by those who want to provide career path to get a specific job, not just what they current qualify for.

6. An individual may have skills developed through experience or self-training that are not reflected in any formal degree. A job relevant task simulation that can test an individual's ability to do a job could be used as an alternative to a degree. If complete or partially complete, it would provide evidence of the person's current skill level. There is a web based system that allows programmers to check out public-domain software. Programmers make changes to the programs and repost them to the site. A prospective employer can look at the quality and sophistication of the changes made by the programmer and see if they want to hire them.

7. Electronic courses like those available at "Khan Academy" provide free content to the general public. However, just providing the content does not provide the help and feedback typically provided by a good teacher or tutor.

8. A quality ITS can assess a student's performance and provide assistance when the student runs into difficulty. It's fairly easy to capture and address commonly ask questions and concept misconceptions. These can be automatically detected and feedback can be provided. Through extended use and crowed sourcing student questions and responses, the system can evolve in the scope of material and problems it can handle.

9. Having an accurate preview of a job is a great predictor of how long someone will stay in the job and their overall job satisfaction.

10. One hurdle to overcome acceptance of freely available education and training courses is the ability to verify that the person is actually doing the work without too much outside help. Some methods need to be in place to verify the completed work.

11. LMSs can be used to track training performance. The Advanced Distributed Learning (ADL) has created the experience application programming interface (xAPI) as a standard to collect and transmit performance data from different training systems and LMSs. Some standardization is required to track a sailor's performance across training systems and throughout their career.

References

Burke, R. (2017, January 24). Changes ahead for Navy personnel system, sailors' training pipeline. Retrieved from https://www.navytimes.com/articles/chief-of-naval-personnel-vice-adm-robert-burke-personnel-system-training-pipeline.

Faram, M. D. (2017, February 19). Get to the fleet faster - Big changes coming to A schools. Retrieved from https://www.navytimes.com/articles/navy-training-overhaul.

CHAPTER 6 – Coordinating Evidence Across Learning Modules Using Digital Badges

Ross Higashi[1], Christian Schunn[1], Vu Nguyen[2], and Scott J. Ososky[3]
University of Pittsburgh[1], Carnegie Mellon University[2], US Army Research Laboratory[3]

Introduction

No matter how successful a learning module or intervention such as an intelligent tutoring system (ITS) is at producing learning, the fruits of those efforts cannot be employed efficiently without a suitable means for representing and conveying which learners possess which skills. Who will know to hire or promote this more knowledgeable individual, if there is no clear sign that they are more accomplished? Digital badges are digital artifacts that function as markers of achievement. Often described as building on the combined traditions found within Scouting (e.g., Boy Scouts or Girl Scouts) and online gaming (e.g., Xbox Live Achievements, PlayStation Network Trophies), badges are issued to an individual when the individual meets specific criteria embedded in program-relevant activities (Ostashewski & Reid, 2015).

Functionally, badges are commonly framed as open digital microcredentials (e.g., Ifenthaler, Bellin-Mularski & Mah, 2016; University of Minnesota, 2017). Openness means that any party should be capable of issuing badges. Digital means that the badges themselves exist in an online environment and are thus amenable to digital transmission, e.g., over the Internet. Finally, the "microcredential" nomenclature emphasizes badges' common purpose with traditional credentials such as diplomas and trade certifications, but with a finer-grain size.

A 2014 survey of an early-adopter cohort of badge developers identified three common design goals for badges: "recognizing learning", "assessing learning", and "motivating learning" (O'Byrne, Schenke, Willis & Hickey, 2015). A growing number of studies (e.g., Abramovich, Schunn & Higashi, 2013; Reid, Paster & Abramovich, 2015; Suhr, 2014) have investigated effectiveness in individual areas, but the juxtaposition of the three is also informative. This is because, as with existing credentials, the assessment, recognition, and motivational components of badges are intertwined: the recognition afforded by a credential depends upon the fair assessment of the skill during the awarding process. Earners may be motivated to gain the credential because it is instrumental for scholastic or career advancement (utility), because it displays their prowess to others (achievement), or perhaps because they see owning the badge as consonant with their personal or professional identities. In all cases, the link between possessing the skillset and acquiring the badge depends, fundamentally, upon the validity and credibility of the assessment process. An attentive evaluator would reject (with prejudice) a badge that claims one thing but measures something else, and no learner would be excited to earn a token thus discredited.

Thus, digital badges could substantially improve the efficiency of skill-based personnel or resource assignment by effectively surfacing learners' skills as they are developed, at a higher frequency and with greater granularity that traditional credentialing processes, although still at grain sizes large enough to be meaningful to outsiders. This may bring with it advantages for learner motivation, and by extension, improve learning outcomes as well. Yet, these things are only possible if the badges contain a valid and credible assessment of the indicated skills. In essence, the entire badging enterprise – and indeed, that of microcredentialing in general – hinges upon the question of why a viewer should believe the badge's claim.

In this chapter, we present a conceptual model for a badge system, illustrated within a computer programming learning environment. The model is built upon theoretical foundations and practical use cases, which are leveraged in order to derive specific design considerations. The chapter concludes with the potential expansion of the badge system and opportunities for future research.

Related Research: Designing Badges for Assessment

Theoretical Framework

To productively connect assessment and evidence, we turn to the evidence-centered assessment design (ECD) framework laid out by Mislevy, Steinberg, and Almond (2003; Mislevy, 2006). Under this framework, an assessment is fundamentally understood to be an argument from evidence, designed for a purpose. An assessment is valid if (and only if) its argument is sound, using observed data – things the student has done – to warrant claims that the student knows certain things. Furthermore, the knowledge claim must be useful toward a real purpose, i.e., relevant to a real decision about a student possessing a skill. In short, a valid assessment makes a claim that an individual knows something, backs that claim with evidence, and leads to a conclusion that is usable for a decision.

How, then, might we design digital badges to embody a valid assessment claim? In the process of design, it is typical to connect these elements in reverse, reasoning from ends to means in a "backward design" process (Wiggins & McTighe, 2011). Intuitively, *purpose* will determine "which features and expectations are central, and which are irrelevant" (Messick, 1994; in Mislevy, Steinberg & Almond, 2003). This means that we must start with the *purpose* of the badge, i.e., why anyone cares whether a student possesses a certain skill in the first place. From there, we can ensure that the badge makes an appropriate claim to fulfill that purpose, and supplies evidence that compellingly backs its claim.

Use Cases

To gain better traction on this issue, we focus on three specific use cases. "Use cases" identify specific, representative scenarios in which the product-under-design must fulfill a certain need, in a certain context. These concrete scenarios allow designers to understand requirements and reference the scenarios as litmus tests for the sufficiency of proposed designs.

Much of the policy interest in badges frames them as credible indicators of knowledge or skills, usable for making decisions about admission to program of study or employment, or for guiding one's own learning (Duncan, 2011; LRNG; MacArthur, 2011). Therefore, we begin by proposing the following three use cases:

- **Use case 1:** Digital badges should help a **college admissions officer** decide whether an applicant possesses *sufficient academic preparation to begin learning college-level content*.
- **Use case 2:** Digital badges should help a **learner, mentor, or ITS** *choose an appropriate next task* or topic for learning.
- **Use case 3:** Digital badges should help an **employer** decide whether an applicant will *be able to perform certain tasks well on the job*.

Aligning Purposes, Claims, and Arguments

Across the three use cases, two distinct kinds of "purposes" have appeared.

Claims about Readiness to Learn. The college admissions officer and student/mentor (use cases 1 and 2) are both interested in the learner's readiness to learn certain new content. The logic implicit in this framing is well established within the learning sciences: learners can only learn certain novel concepts after certain

prior learning has put them within reach of it. This type of claim is typically most relevant in cases of formative assessment – that which is intended to inform mid-course adjustments in learning trajectories.

Research related to the zone of proximal development (Chaiklin, 2003; Vygotsky 1933), conceptual change (DiSessa & Sherin, 1998), learning hierarchies (Duncan & Hmelo, 2009), and effective tutoring (Koedinger, Corbett & Perfetti, 2012; Wood, Bruner & Ross, 1976) has unpacked theoretical and practical concerns around this phenomenon. For our purposes, this distinction is particularly important because the readiness-to-learn *purpose* informs the *kind of assessment argument* that is needed to support it. Specifically, the argument must allow us to conclude that the student possesses knowledge *X* in such a way that it has prepared them to learn *X'*.

Claims about Proficient Reapplication. The importance of alignment is also evident when we consider the second major *purpose* contained in our use cases: that of the hiring manager (use case 3), who may be primarily interested in whether a job applicant will be able to perform certain skilled job functions reliably once hired. That is, the hiring manager wants to know that the applicant will be able to apply the skill proficiently and appropriately under working conditions. This type of claim is commonly associated with summative assessment – that which is intended to summarize qualifications.

The logic underlying this framing is of a wholly different nature: it concerns the *transfer* of learning from the context in which it was learned (e.g., through a lecture in the classroom or in a training environment) to new contexts in which it should be applied (e.g., to a problem-solving task in the workplace). Learning scientists have given extensive attention to the fact that two students who appear to understand something well, may still differ in their ability to "transfer" that knowledge to a new situation. This is often a concern for intelligent tutors or simulation environments in which the learning environment is systematically simplified to make it digital. Many models have been proposed and tested of both the underlying causes and ways in which transfer performance might be improved (e.g., Hammer, Elby, Scherr & Redish 2005; Kirschner, Sweller & Clark, 2006). An assessment argument for this purpose will likewise need to be of an entirely different character from a readiness-to-learn argument. Rather than focusing on a student's ability to comprehend future material, this purpose demands that the assessment argument be made about the student's ability to transfer knowledge to the work context.

Design Takeaways

In practical terms, our digital badge designs must make the assessment claims that speak to both the "readiness to learn" and "reapplication" *purposes*, and back those claims with evidence. Stated in "forward" order, digital badges make assessment claims of the form: "Using past performances as evidence, we assert that the earner of this badge possesses the indicated skill, and will be able to apply it appropriately and build upon it in the future".

With these general design objectives in place around the alignment of *assessment claims* to *assessment purposes*, we now turn to the specifics of the "evidence subsystem" that facilitates the formation and delivery of an appropriately "backed" assessment argument.

Design Principles for a Badge-Based Evidence Subsystem

In this section, we propose detailed principles for the design of *evidence subsystems* within broader badging arrangements, designed to support valid arguments-from-evidence about the knowledge and skills of badge earners. Our objectives are twofold: 1) to more specifically address concerns about *establishing evidentiary warrant* and 2) to provide one concrete, practical solution.

Design Principles

As a design progresses from "clarified problem statement" to "specific solution", it sometimes acquires features that might be considered idiosyncratic and do not further any design goals. To make a small but powerful set of implementation decisions, while avoiding superfluous prescriptions altogether, we rely on a small set of guiding principles to help us maintain design discipline around the complex notion of "evidentiary warrant with fidelity to purpose". These principles reflect, in a sense, the two complementary facets of a parsimonious design: 1) the provided evidence must be *sufficient* to establish (or "warrant") the claim, and yet 2) are no more complex than *necessary* due to practical and logistic concerns. In both cases, we draw again on our use cases, as both necessity and sufficiency are defined relative to *purpose*.

Principle 1: Evidentiary Strength

Strength of evidentiary warrant increases with both quantity and diversity of evidence. To address the issue of sufficiency of evidence, we draw upon the intuition of the replication study, or triangulation. In the sciences, we recognize that no single experiment, study, or perhaps even theory provides a complete picture. The single datum is thought of as fundamentally impoverished, lacking a robustness of perspective that can only be established through consideration from multiple angles. A conclusion supported by only a single data collection and analysis is at best promising, but provisional. Analogously, an assessment statement is only weakly warranted by a single piece of evidence.

Furthermore, while additional evidence of the same type would increase our certainty in the conclusion somewhat, it does not achieve the same effect as a concurrent result from a fundamentally different analysis since the inherent weakness of any single source is maintained in an exact replication. Real replication is not duplication of an analysis, but the reproduction of an equivalent result in a different context.

We extrapolate three important design features from this principle. First, we recognize that evidentiary warrant is not dichotomous, but dimensional — a small amount of evidence would provide weak support for the assessment claim, while more or better evidence would provide stronger support. It is not all-or-none. Second, we recognize the existence of multiple, qualitatively different types of evidence that might speak to the validity of the skill or knowledge claim. This is concurrent with a central theme of modern learning science research: in addition to the classical cognitive theories of conceptual understanding (typically assessed by, e.g., standardized multiple-choice exam), several strands of research focus on environmental and social factors involved in promoting the development of transferrable skills (e.g., Gresalfi, 2009; Lave & Wenger, 1991; Lee, 2008). Finally, as a combination of the first two, we recognize that the best conceptualization of the evidence space is in fact multidimensional: different amounts of different kinds of evidence.

We conclude with two design decisions based on the aforementioned design features:

- **Design Decision 1:** The framework must support the inclusion of multiple kinds of evidence concurrently. The badge system must, at a technical level, support the inclusion of evidence types beyond traditional exam scores. For instance, a portfolio of works completed by the student may be considered a valid form of triangulating evidence. In the ECD framework, each piece of evidence is free to rely on its own theoretical sense of internal validity, as long as it is valid according to that standard.

- **Design Decision 2:** The design must recognize and represent multiple "levels" of evidentiary warrant. These should be tied to the quantity and diversity of evidence provided. Strength of warrant must be kept meaningfully separate from the level of mastery being claimed.

Principle 2: Evidentiary Necessity

The strength of evidence needed to support a claim is based on the weight of the decision that will be made. Stronger warrant is needed for higher-stakes decisions. Having established that the strength of evidentiary warrant can vary with the amount and diversity of the evidence, the second issue we must address is, "how much evidence is enough?" Since there is no definitive scientific answer to this question (Popper, 2005), we borrow an intuition from legal theory, where the use of imperfect evidence to justify conclusions is common practice. This intuition is that of the sliding "standard of proof", in which the required strength of evidence is higher when the potential consequences are greater (e.g., in criminal vs. civil court).

While the circumstances of a badge evidence evaluation are certainly not the same as being in court, decisions made using badges do vary in terms of potential impact. A student's decision to move on to the next chapter (based on a badge saying they understood the previous concept) is fairly low-stakes and can be made based on even relatively thin evidence because even in the worst case, a learner need only backtrack to review. On the other hand, a college admissions officer is making a substantially higher-stakes decision when using badges to determine an applicant's readiness to learn in college and should be bound to accept only evidence that is more firmly established. Thus, we extrapolate one important design feature from our consideration of evidentiary necessity: our design must indicate in some meaningful way the strength of evidentiary warrant it provides, so that a viewer can have some sense of how firmly its claim should be considered established.

- **Addendum to Design Decision 2 ("2b"):** The design must state or strongly suggest a clear relationship between stakes-appropriateness and strength of evidentiary warrant. This is functionally an addition to Design Decision 2 from the previous section.

Principle 3: Efficiency

There is a practical upper bound on how much evidence an evaluator is able and willing to examine. The more efficient the presentation, the more evidence can be used. Finally, we look at a practical factor that has implications for how much evidence we can effectively (rather than theoretically) bring to bear in backing our assessment argument. Simply put, no evaluator has time to interrogate all the evidence in detail, for every badge that is presented. Looking back to our use cases once more, it is the highest-stakes decision makers – the college admissions officer and the potential employer – who are also tasked with evaluating the largest quantity of badged claims. Yet, per our second principle, these are the decision makers who must consider the evidentiary claims most carefully. Thankfully, good communication design techniques can mitigate this information bottleneck by distilling complex and ungainly information into easier-to-understand visual summaries that can be easily reviewed without loss of the "big picture" regarding evidentiary warrant.

- **Design Decision 3:** Each type of evidence in the badge must be summarizable for quick viewing. The composite strength of the badged evidence claim should also be easily summarized.

This decision will probably not manifest a single large feature, but rather become a criterion in the design of many small features (for instance, whether we choose a single-number vs. a long-list display for certain evidence types).

A Conceptual Prototype of a Badge-Based Evidence Subsystem

Based on the three design decisions we laid out in the previous step, we now present a set of badge evidence subsystem design concepts that implement those decisions. For an illustrative example, we follow the hypothetical case of a Loops Programming badge in an introductory computer science learning context, as it attempts to make the argument that the earner possesses the (relatively basic) programming skills needed to "Use loops to repeat sequences of commands" and "Use conditions to end a loop at the appropriate time". The purpose of this section is not to present an optimal design solution, but rather to illustrate and work through an additional layer of details. We thus frame the framework described in Figure 1 as a "conceptual prototype".

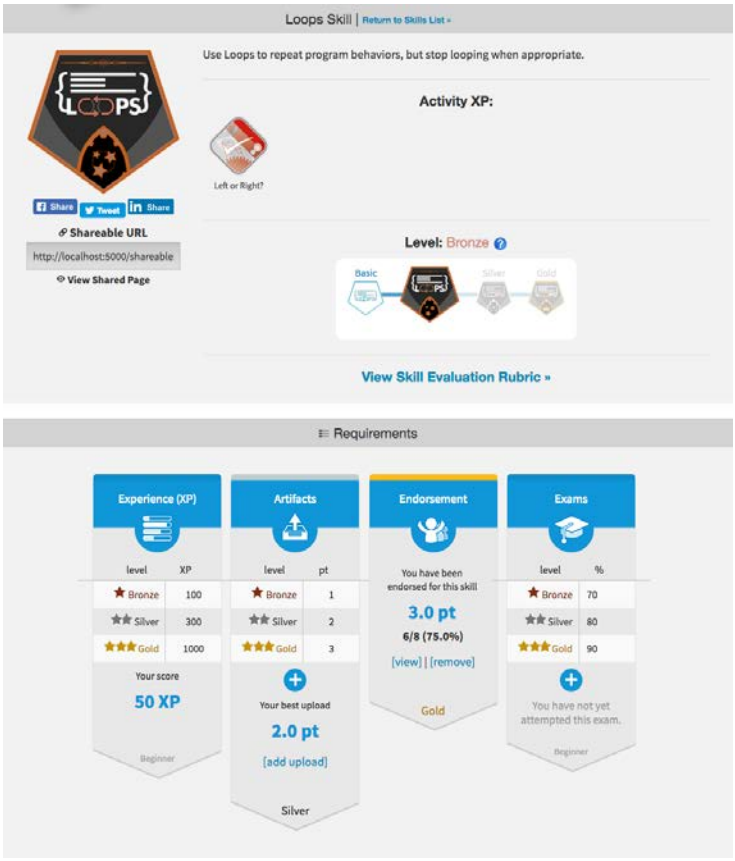


Figure 1. A composite overview of the proposed conceptual prototype.

Four Dimensions of Evidence

Based on Decision 1, to implement a multidimensional representation of evidence based around the types of evidence and amount of each, we have formulated a set of four major evidence categories. These categories are selected based on a combination of practical concerns, and the principle that greater epistemological diversity of evidences provides better triangulation.

XP: Experiences and Experience

The XP category captures the amount of relevant experience in completing skill-relevant tasks that the learner possesses, and represents it summarily as an experience point (XP) total (drawing on a popular video gaming convention).

Our Loops Programming learner might earn 10 XP for completing an online learning module about loops, 15 XP for checking in at a hands-on programming workshop event, and another 25 XP for competing in a robotics programming competition (plus perhaps some bonus points for placing well in the competition). These would be pooled into a single summative "XP" statistic (Figure 2).

Figure 2. An example of the XP tracker for a hypothetical learner's Loops Programming Badge.

Evidence of this type implicitly makes the claim that the badge holder has engaged with (and succeeded at) skill-relevant activities. This connects epistemologically to the main assessment claim that "the badge earner has X skill" largely through associationist theories of conditioning (training the brain to produce the right responses to task-relevant stimuli), rehearsal (repeated exposure to the correct problem-solution pair improves ability to recall the correct course of action in the future), and the motivational theory of behavioral engagement (higher levels of participation in school activities predict higher learning outcomes). These theories suggest that greater levels of experience predict greater proficiency of application, application under diverse contexts, and ability to build on these experiences to facilitate future learning. Summarily, the more one participates in (and eventually completes) skill-related activities, the better one is expected to become at them.

As a metric that lends itself to quantifiability, it is important to determine a rough scale for XP – that is, to set some common expectations around "what 100 XP means". Since we expect XP to relate to the same kinds of outcomes as rehearsal and behavioral engagement, we will use those outcomes to establish an outcome scale. Specifically, higher XP should correspond to greater ease of recall, greater future participation, and greater spontaneous recall. We therefore define the XP reference scale in terms of these quantities (albeit with initially arbitrary cutoffs that can be refined with testing):

- A learner at the 100 XP mark should be able to apply the skill with prompting and assistance.

- A learner at the 300 XP mark should be able to apply the skill with minimal guidance, when prompted.

- A learner at the 1,000 XP mark should be able to apply the skill fluently and spontaneously in novel situations.

Strengths of the XP metric are its easy quantifiability (number of XP) and summarizability (as a single number). When skill-relevant activities can be identified in advance, XP tracking is easily automated. Such systems might even incorporate, e.g., diminishing point returns on highly scaffolded activities for learners who already have high levels of XP. The XP mechanic also allows the incorporation of badges from other badge systems as a form of evidence, as many settings "badge" activity completion. There are two primary weaknesses of XP as a form of evidence. The first is opacity as to exactly what kinds of activity the student engaged in, heavily favoring quantity over quality. The second is opacity of methods for completing tasks; often tasks can be completed through both use of the indicated skill and alternative brute-force methods (e.g., guess-and-check or asking for outside assistance, which learners do for many reasons; see Baker, et al., 2008).

Finally, it is worth noting that interpretability of "XP" leans heavily upon the framing of the badge as a whole. Fluency in "loops programming" is easier to characterize and understand than fluency in "organizational leadership", partly because the latter is so broad. This comparison also reiterates that point values are always relative to badge scope; a task worth many points toward a narrowly defined skillset would be worth only a few toward a broader one.

Artifacts: Contextualized Work Product Examples

The artifacts category captures specific concrete instances of submitted student work. Students describe and self-rate their submissions according to a badge-attached rubric upon submission. This rubric calls attention to salient features that demonstrate the skill, for both students and evaluators. It also serves as a first-order filter for quality and relevance of the submission to that skill (i.e., a student who is giving low ratings to their submission across all categories should recognize it as a poor fit for evidence toward that skill). Evaluators can subsequently interrogate the self-ratings along with the artifact as an estimate of a student's understanding of the skill itself.

A Loops Programming badge-seeker might submit an annotated copy of the source code from a class project as an artifact, and rate that work on the rubricized dimensions of "Using loops to repeat sequences of commands" and "Using conditions to end a loop at the appropriate time". Both the artifact and the rubricized rating would be made available under the "artifacts" evidence category (Figure 3).

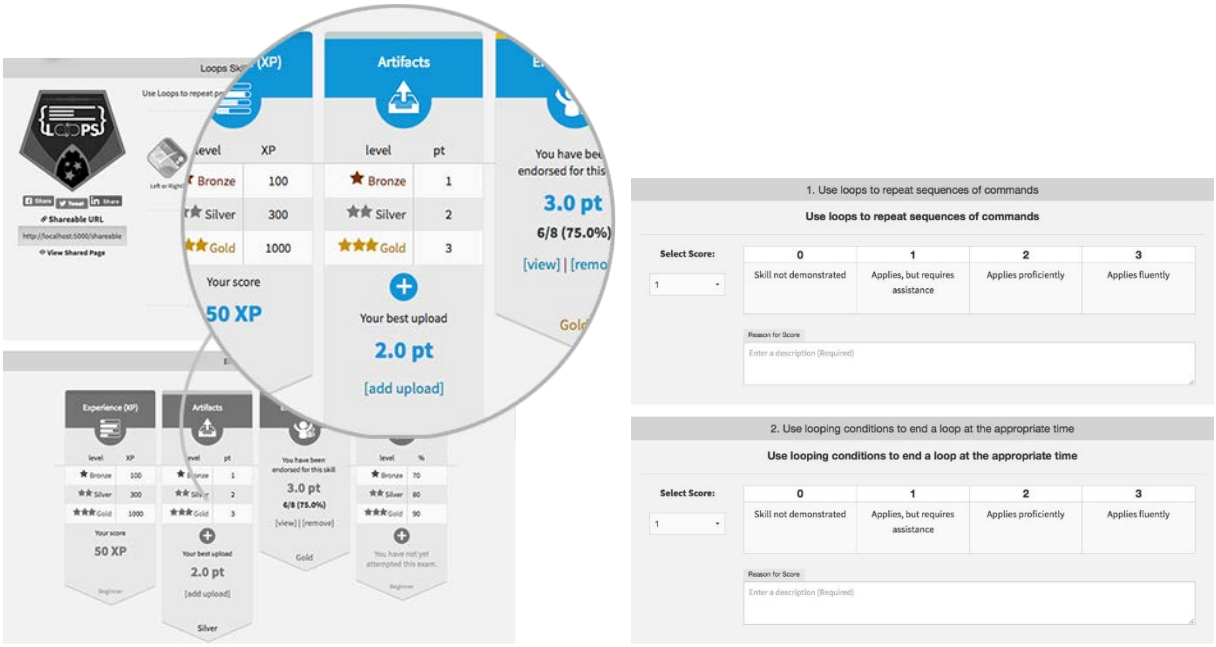


Figure 3. An example of the artifact evidence interface and rubric summary display for the Loops Programming Badge. Clicking the link gives the viewer access to the uploaded file (in this case a piece of source code), a copy of which is permanently stored on the server.

Epistemologically, evidence of this type draws on the same assessment traditions as portfolio assessment: the objects in the portfolio directly demonstrate advanced examples of skill, and a well-constructed rubric maps relevant features of the skill to relevant qualities of the objects. The rubric-rated work product allows scoring of skill-relevant performance in an authentic context.

The validity argument from this data is connected to 1) the transparency of skills in submitted objects; 2) the rubric's relevance in representing the "right" qualities of the knowledge or skill for rating for submitted objects; and 3) the reliability and trustworthiness of the rater to give accurate ratings. The strengths of this approach lie in its dual provision of a work product with its context intact, alongside a scoring system mapped directly onto the skill claim being made. It further allows evaluators to inspect the quality of the work by examining the artifact itself. These speak directly to the earner's ability to apply and reapply the skill proficiently. The primary weaknesses of this evidence type are its vulnerability to unreliable self-raters, dependence upon the expertise of the person viewing the badge evidence (i.e., non-experts will not be able to understand the evidence), and the time-consuming nature of inspecting the artifact itself, especially for more complex artifacts (i.e., few people will take the time to investigate details for objects with a large amount of detail, such as thousands of lines of code).

Endorsement: Expert and Participant Verification

The endorsement category captures assessments of intangible expertise conferred by experts and peers within a practice space. Teachers, mentors, and peers are solicited to provide a short online endorsement of the student's proficiency at the given skill, including both a rubricized evaluation and a short written statement. This replicates, to some extent, the recommendation-writing system embedded in both college and job application processes. Judgments can be based in the artifacts within the badging system (i.e., adding an outside evaluation of the same submitted object) or based in observation of behaviors or discussion (i.e., connected to very different sources of evidence).

The Loops Programming badge earner might solicit an endorsement from the professional programming mentor on her robotics team. The mentor would log in to the online system and rate the student on the rubric for the Loops Programming skill ("Using loops to repeat sequences of commands" and "Using conditions to end a loop at the appropriate time"). If the mentor has registered correctly within the system, the endorsement is marked as being "by an expert" (Figure 4).

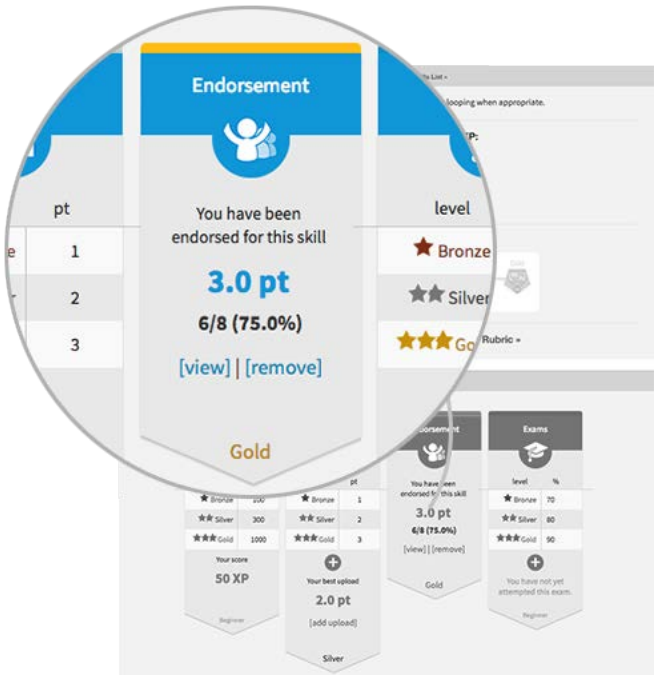


Figure 4. An example of the endorsements this learner has received for Loops Programming. The rater was identified as an expert based on credentials within the system.

Evidence of this type contains two features with different epistemological roots: the rubric and social sources of endorsement. The use of the rubric maps scoring onto evaluation of relevant features of the skill, as it did in the artifact category (the rubrics should be equivalent for this reason). The scoring process for the rubric and the provision of the written recommendation, however, tap sociocultural mechanisms for evaluating skill.

Epistemologically, there are certain kinds of skill that are notoriously hard to evaluate; tacit knowledge, for instance, refers to the unspoken and generally unmeasurable expertise that certain individuals have for making good decisions under poor information conditions (Collins, 2001). Army officers making difficult discipline and personnel decisions display a remarkable degree of similarity and expert consensus, but the underlying skill is exceedingly difficult to measure through means other than agreement among those experts (Schunn, McGregor & Saner, 2005; Sternberg & Horvath, 1999). In sociocultural frameworks such as Situated Learning (Lave & Wenger, 1991), acceptance by the community of practice leading to increased participation *is* the mechanism by which expertise is gained.

The major strength of this evidence category is in capturing intangible, yet historically reliable, assessments of expertise by knowledgeable others within the learning space. This type of evidence is also uniquely positioned to reflect on certain types of skills that are impossible to evaluate through other means, such as collaboration quality, which is inherently social. The primary weakness of this approach is that it is difficult to establish the legitimacy of a rater, without the argument quickly becoming circular. This is particularly

difficult when dealing with "non-expert" raters, e.g., peers, whose actual level of expertise (and hence reliability) might be quite low. We report this relationship (e.g., "Expert" or "Peer") in the interface alongside the endorsement, so that an evaluator can choose to take this into account.

Exams: Transfer to Strategically Chosen Tasks

The exams category captures learner performance on designed measurement tasks. Implicitly, this tests students' ability to transfer their learning to the testing context. Students take a scored exam of some sort – most commonly, this will be of a traditional examination format, in which students respond to crafted prompts designed to measure one or more validated skills. In the case of exams that measure multiple skills, only the items relevant to the badged skill would count toward this score.

The Loops Programming badge might accept and electronically import exam scores from a list of trusted sources, such as vetted online courses and AP exams. These data providers would report a loops-relevant subscore, or scores of exams wholly relevant to the topic of loops programming (e.g., a Loops Chapter Exam). The Loops Programming badge interface displays the list of acceptable exams and provides a link to at least one free online exam option, as well as the highest available score among the eligible exams (Figure 5).

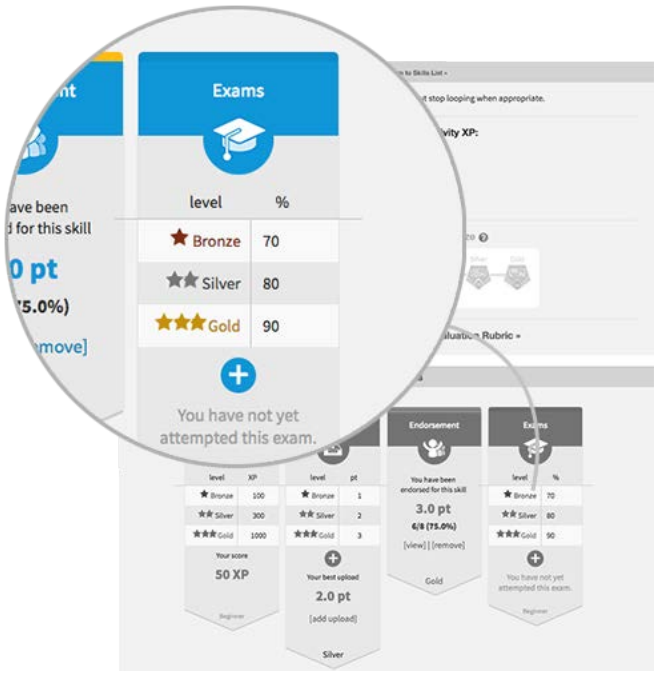


Figure 5. An example of the exam evidence summary for the learner's Loops Programming Badge. If the exam has a viewable sample item/page, clicking the link opens it for inspection.

Epistemologically, evidence of this type draws on whatever measurement techniques are "baked in" to the exams themselves. Typically, these draw upon cognitive theories of skill development and transfer ("someone who knows X will answer A for test question item I"), backed by traditional methods used to validate such items, such as Item-Response Theory. ITSs often use evidence of this type.

The main strength of this approach is its ability to incorporate traditional measurement media as a form of evidence. They are as valid as their own methodological backing, and lend that validity to the badged evidence pool. They also allow for developing individual items that target each aspect of a large skill to allow

for "complete" coverage of the skill, and they encourage learners to completely master the domain rather than focusing on areas of interest. Finally, they also carry with them the cultural and policy "weight" of these exams as they are used in the world today – insofar as the chapter exam is trusted as a measure of knowledge or skills today, adding it as evidence to a badge increases the badge's evidentiary warrant toward that knowledge or skill.

The main disadvantages of this evidence type are logistical: finding test items that tap a given skill and disentangling a skill-specific score from a larger exam is difficult to scale as a general practice. In many cases, standalone exams are not well validated, capturing only superficial aspects or only declarative knowledge related to the skill (e.g., memorized answers to familiar questions), even though they may be accepted in practice.

As a final note, some exams may allow only a limited number of attempts to prevent users from seeking high scores by "brute force" retaking of the test. The exams evidence construct is agnostic to this choice, leaving the decision to the individual exam administrators so as not to impinge their authenticity. While this creates the possibility that a user could be stuck with an irrevocably low score on a given test due to its strict retake policy, the user's ability to select from among multiple exams for the exam evidence allows a workaround (albeit perhaps using a more onerous or less prestigious exam).

Representing "Composite Claim Quality"

Decision 2 (with addendum 2b) in the previous section task us with designing a system element that recognizes and represents a sliding scale of evidentiary warrant, and simultaneously relates strength of warrant to stakes of decisions.

Intuitively, we want to pick a single design element to represent both *strength of evidence* and *suitability for higher-stakes decisions*. We want to pick some visual device that we can use to represent higher versus lower levels of this shared "claim quality" dimension. However, in doing so, we should also consider other "qualities" of a badge claim that users might confuse it with. Two in particular seem to be likely sources of misinterpretation: one quality that badges frequently indicate is "more advanced skillsets". This dimension is often represented by badge material (e.g., bronze vs. silver vs. gold), badge size, or cumulative marks such as stars. A second quality is "more advanced proficiency within the skillset" (e.g., very high levels of proficiency with beginner-level coding techniques).

The first potential confound is neatly separable. "Proficiency at more advanced versions of the skill" is better represented by a different claim, as it represents different knowledge and techniques that should be separately represented and assessed.

The other – more advanced proficiency within the skillset – maps nicely onto an evidentiary dimension we have already identified: it is the same as the "level of proficiency" that our multidimensional framework represents as rubric and XP scores. Furthermore, it correlates well with the notion of "higher stakes" – decisions such as hiring and college admissions are likely to want higher proficiency levels as well as more evidentiary certainty. Conversely, a claim of a higher level of skill deserves to be inspected more carefully and backed by stronger evidence, especially since knowledge of "all" of a skillset is likely to need a greater quantity of evidence to establish simply due to the larger "surface area" of the knowledge being claimed. Thus, we include this third dimension of "level of proficiency claimed" as a third dimension sharing the same design feature as the first two. Our composite quality to be represented is now composed jointly of "strength of evidentiary warrant", "suitability for higher-stakes decisions", and "proficiency level claimed". For our design, we elect to use badge levels of bronze, silver, and gold to indicate the composite quality of *stronger assessment argument*, *suitability for higher-stakes decisions*, and *higher levels of proficiency being claimed*.

This is somewhat complicated by the fact that not all types of evidence may be available or suitable for all types of skill or knowledge claims. For instance, a "collaboration" skill will probably not involve an exam. This means that such a skill has, at most, three possible types of evidence. Other skills may not have reasonably presentable work products (perhaps projects are too large, too small, or too confidential). Some may not have quantifiable "experiences" because they are innumerable, pervasive, or untrackable. Learning done by solo learners online may not have anyone to endorse them. In short, our system must be robust to a number of potential evidence types that varies anywhere from 1 to 4.

We therefore define our "composite quality" index to include the following levels, using relative counts of evidences, rather than absolute counts:

- A **Gold badge** provides **all possible evidence types**, with scores of **80% or higher** on all available rubrics and exams.
- A **Silver badge** provides **all but one of the possible evidence types**, with scores of **70% or higher** on all included rubrics and exams.
- A **Bronze badge** provides **at least one type of evidence**, with a scores of **60% or higher** if it is a rubric or exam.
- "Corner cases": For badges which permit only two types of evidence, Silver is omitted. Badges that have only one evidence type available are referred to as "binary badges", and they can only be earned or unearned.

Recommendations and Future Research

In this chapter, we have laid out a design blueprint for an assessment system rooted in the provision of evidence in and through digital badges. We began with theoretical foundations and practical use cases, and from these derived design principles, then a conceptual prototype. There remains work to be done in implementing the prototype design. For instance, the user interface design of the system plays a key role in framing and explaining the evidence requirement and submission system to both badge earners and viewers. Given the complexity and novelty of the proposed arrangement compared to conventional "one and done" exams and certification routines, completing the full interface is no mean feat. An adequately scalable technology platform would also be required to host both the evidence-collecting activities (e.g., endorsing) and the resulting badge data indefinitely.

At an organizational level, there is the challenge of identifying "expert" individuals who should appear as such when endorsing learners. Would such a qualification be imported through an existing registry of sorts (e.g., national teacher councils), or would they be "bootstrapped" into the system by some form of qualification exam? There is also a short-term need to either develop in-house exams for skills, or vet existing exams to determine which ones can qualify as exam evidence, followed by the need to establish a digital data pathway for importing learners' scores and associating them with the correct users in both systems. This activity extends into a long-term need to establish partnerships with commercial testing services and perhaps even individual states or school systems.

With these things in place, however, additional opportunities also open up. The digital nature of badges and of the evidences allows for high levels of integration into, e.g., intelligent tutoring or game-based systems which can automatically award XP, upload student work, and prompt teachers to provide endorsement at opportune moments in the learning process (for an expanded overview of this topic, see Ososky, 2015). Social media integration could enable additional ways of acquiring evidence; perhaps community-based

"popular" endorsement by large numbers of peers could provide an alternative to "expert" endorsement for certain skills.

These are only a few of the assessment opportunities that could be tapped with digital badges. Nevertheless, future sources of evidence, however creative, will be efficacious only if these diverse and powerful sources of evidence are harnessed through a robust framework for assessment. In this chapter, we have advanced the digital badge conversation toward the development and adoption of a principled assessment framework, for it is only with this in hand that digital badges can truly unlock their potential to inspire and reward learning.

References

Abramovich, S., Schunn, C. & Higashi, R. M. (2013). Are badges useful in education?: it depends upon the type of badge and expertise of learner. *Educational Technology Research and Development*, *61*, 217–232.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A. & Koedinger, K. (2008). Why students engage in" gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185–224.

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context*, *1*, 39–64.

Collins, H. M. (2001). What is tacit knowledge. In T. R. Schatzki, K. Knorr Cetina & E. Von Savigny (Eds.), *The practice turn in contemporary theory*, p107-119. New York, Routledge.

DiSessa, A. A. & Sherin, B. L. (1998). What changes in conceptual change?. *International journal of science education*, *20*(10), 1155–1191.

Duncan, A. (2011). Digital badges for learning. Speech given at *4th Annual Launch of the MacArthur Foundation Digital Media and Lifelong Learning Competition*. September 15, 2011.

Duncan, R. G. & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching*, *46*(6), 606–609.

Gresalfi, M. S. (2009). Taking up opportunities to learn: Constructing dispositions in mathematics classrooms. *The Journal of the Learning Sciences*, *18*(3), 327–369.

Hammer, D., Elby, A., Scherr, R. E. & Redish, E. F. (2005). Resources, framing, and transfer. *Transfer of learning from a modern multidisciplinary perspective*, 89–120.

Ifenthaler, D., Bellin-Mularski, N. & Mah, D. (Eds.). (2016). *Foundation of Digital Badges and Micro-Credentials: Demonstrating and Recognizing Knowledge and Competencies*. Switzerland: Springer International Publishing. doi:10.1007/978-3-319-15425-1

Kirschner, P. A., Sweller, J. & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational psychologist*, *41*(2), 75–86.

Koedinger, K. R., Corbett, A. T. & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, *36*(5), 757–798.

Lave, J. & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge university press.

Lee, C. D. (2008). Cultural modeling as an opportunity to learn: Making problem solving explicit in culturally robust classrooms and implications for assessment. *Assessment, equity, and opportunity to learn*, 136-169.

LRNG. (n.d.). Products: LRNG. Retrieved September 4, 2016, from http://about.lrng.org/products/.

MacArthur Foundation. (2011, September 15). Digital Media & Learning Competition Provides $2 Million for Innovations in Digital Badges. Retrieved September 7, 2015, from https://www.macfound.org/press/press-releases/digital-media-learning-competition-provides-2-million-for-innovations-in-digital-badges/.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *32*(2), 13–23.

Mislevy, R. J. (2006). Issues of Structure and Issues of Scale in Assessment from a Situative/Sociocultural Perspective. CSE Technical Report 668. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Mislevy, R. J., Steinberg, L. S. & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, *1*(1), 3–62.

O'Byrne, W. I., Schenke, K., Willis III, J. E. & Hickey, D. T. (2015). Digital Badges Recognizing, Assessing, and Motivating Learners In and Out of School Contexts. *Journal of Adolescent & Adult Literacy*, *6*(58), 451–454.

Ostashewski, N. & Reid, D. (2015). A History and Frameworks of Digital Badges in Education. In *Gamification in Education and Business* (pp. 187–200). Springer International Publishing.

Ososky, S. (2015). Opportunities and Risks for Game-Inspired Design of Adaptive Instructional Systems. In D.D. Schmorrow & M.C. Fidopiastis (Eds.), *Foundations of Augmented Cognition: 9th International Conference, AC 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings* (pp. 640–651): Springer International Publishing.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Reid, A. J., Paster, D. & Abramovich, S. (2015). Digital badges in undergraduate composition courses: effects on intrinsic motivation. *Journal of Computers in Education*, *2*(4), 377–398.

Schunn, C. D., McGregor, M. U. & Saner, L. D. (2005). Expertise in ill-defined problem-solving domains as effective strategy use. *Memory & cognition*, *33*(8), 1377–1387.

Sternberg, R. J. & Horvath, J. A. (Eds.). (1999). *Tacit knowledge in professional practice: Researcher and practitioner perspectives*. Psychology Press.

Suhr, H. C. (2014). *Evaluation and Credentialing in Digital Music Communities: Benefits and Challenges for Learning and Assessment*. MIT Press.

University of Minnesota. (2017) Open Microcredentials | Hype Cycle for Education. Retrieved January 15, 2017, from http://hypecycle.umn.edu/hype-cycle-technologies/open-microcredentials.

Wiggins, G. & McTighe, J. (2011). What is backward design? *Understanding by Design*, 7–19.

Wood, D., Bruner, J. S. & Ross, G. (1976). The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, *17*(2), 89–100.

CHAPTER 7 – Leveraging Domain Models for Personalizing Problem Solving and Learning

Louise Yarnall[1], Eric Snow[1], Erica Snow[1], and Irvin R. Katz[2]
SRI International[1], Educational Testing Service[2]

Introduction

This chapter explains the basic elements of one method for specifying assessment tasks—evidence-centered design (ECD)—and describes the application of the approach to the complex skills of computational thinking (CT) and science inquiry. Both are seen as useful for improving students' capacity to perform the types of flexible applied reasoning and problem-solving skills that are in growing demand in technology-rich workplaces. The chapter closes with a discussion of the implications for content authoring and assessment of problem solving tasks in a Generalized Intelligent Framework for Tutoring (GIFT) system.

Domain Modeling in Science Inquiry and Computational Thinking

To improve American students' readiness to engage in the flexible problem solving demanded by technology-rich workplaces, policy leaders have called for greater instruction and assessment in applied reasoning and problem solving. For example, the Next Generation Science Standards (NGSS)—which have been reviewed by 40 states and formally adopted by 18 as of 2016—embody American science education's shift away from a longstanding emphasis on declarative and conceptual knowledge to a greater emphasis on applied knowledge and skill (NGSS, 2009). Also, the White House has called for more than $4.2 billion to develop instruction to engage younger learners in computer science, an applied science field noted for its emphasis on reasoning and problem solving (Smith, 2016). Acknowledging the challenges of both teaching and assessing such applied skills, the US Education Secretary has emphasized the utility of automated assessments in the nation's standardized testing system (US Department of Education, 2015). Some educators developing assessments and instruction for these science reform initiatives have employed the standardized modeling approaches of ECD (Messick, 1994; Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006). The goal of this chapter is to consider, based on an exploration of two cases from science, technology, engineering and math (STEM) education reform, the capacity of ECD to improve the capacity of developers of intelligent tutoring systems (ITSs) to apply their automated methods to a broader range of applied reasoning instructional problems.

This chapter begins with a discussion of the need for modeling methods to support instruction around applied reasoning and problem solving. This discussion covers the challenges that ITS educational designers face to engineer learning processes around such skills consistently and rigorously, and the specific challenges around teaching applied reasoning skills in two fields of high interest in education policy—CT and science inquiry. Once these needs are described, the chapter then focuses on modeling methods. This section discusses how ECD modeling methods have been used to analyze complex reasoning processes in CT and science inquiry, and then presents partial examples of ECD models in each field. The discussion around these two examples will bring out the common ECD features, but also illustrate how to tailor ECD modeling to focus on assessment at one level of proficiency (the CT case) or assessment at increasing levels of proficiency (the inquiry science case). The chapter concludes by considering the potential of using ECD modeling methods more broadly to inform the design and development of standardized ITS content templates for applied reasoning and problem-solving skills in the GIFT system (Sottilare, Graesser, Hu & Holden, 2013).

Background on ITS Designs to Support Flexible Problem Solving Skills

To introduce this discussion, it is helpful to review how ITS developers model the path of learning and instruction. At the most basic level, ITS developers begin by specifying two models: s *domain model* of the knowledge to be learned and a *student model* characterizing the types of variations in learner's characteristics that affect learning progress, such as knowledge level, affect, and motivation (Park & Lee, 2003). Domain models often are derived from analyses of learning science and expert knowledge. Student model specifications may be based on common student misconceptions and errors and tutoring methods. In practice, ITS systems check learner performance against these two models, continually updating estimates of learner competency, and then making individualized learning recommendations (see reviews of ITS modeling approaches by Desmarais & Baker, 2012; Galyardt, 2015; Koedinger, Brunskill, Baker, McLaughlin & Stamper, 2013).

In the latter 20th century, ITS modeling began in domains with more regimented procedural and conceptual structure, such as mathematics and computer programming, but in recent years, ITS developers have focused more on modeling learning in domains with more learner-driven procedural and conceptual structure, such as writing and social sciences. They also have begun modeling the more dynamic facets of learning, such as perseverance, motivation, and affect.

One of the findings from this more recent ITS research is that learning to solve problems in less structured domains relies to a greater extent on learners' *self-regulated learning processes*, which refers broadly to the range of skills involved in goal-setting, planning, and metacognition (Butler and Winne 1995; Pintrich 2000; Winne 2001; Winne and Perry 2000; Zimmerman 2000, 2001). To guide modeling of self-regulated learning processes in these domains, ITS designers rely on learning science and empirical research. Using this information, ITS designers specify the learner performance data to gather and the prompts to provide, such as those that human tutors offer. Some of the data they gather focuses on cognitive aspects of learning and some focuses on motivational and behavioral aspects of learning. Both present complex modeling challenges. The focus of this chapter is primarily on the challenge of modeling both cognitive and metacognitive aspects of learning applied reasoning.

To provide a broad conceptual framing of the challenge of modeling applied reasoning, Jonassen (2000) provides a helpful guide. He was an early theorist who considered how to define both the cognitive and metacognitive aspects of applied problem-solving processes in multiple domains. He developed a problem-solving design framework that characterizes the range of problem solving tasks, from simple to complex. His framework honed in on three distinct forms of knowledge and skill that individuals need to excel in problem solving: 1) knowing *when* to apply specific principles and concepts; 2) knowing *what solutions* are desired based on local beliefs, needs, and evaluation criteria; and, 3) developing skills of *making decisions* about solution procedures (e.g., technical standards of elegance and parsimony, practical cost-benefit considerations). These forms of knowledge not only depend on accurate recall of factual, conceptual, and procedural knowledge, but they take into account the dynamic factors that influence how that knowledge is applied: timing, situation, and judgment.

As is seen in the next sections of this chapter, these fundamental considerations defined by Jonassen offer a useful introduction to understanding the challenges of teaching and learning applied reasoning and problem-solving knowledge and skills.

CT: Importance and Construct Definition

Over the past decade, there has been a growing focus on the concept of CT both within and outside computer science (e.g., Adams, 2008; Astrachan, Hambrusch, Peckham & Settle, 2009; Denning, 2009; NRC, 2010,

2012; Wing, 2006, 2008). From this early work, we can broadly view CT as the intellectual and reasoning skills needed to master a range of reasoning skills, such as algorithmic thinking, pattern recognition, abstraction, decomposition, and other computational techniques to problems in a wide range of fields. CT is distinct from the specific skill of programming in a particular language; rather CT serves as a framework that captures a range of concepts, dispositions, and applied reasoning practices used when solving problems in many domains.

To further specify the range of knowledge, skills and attributes underlying CT, Brennan and Resnick (2012) identified three main components: concepts, practices and perspectives. CT concepts include the key ideas and knowledge that are central to computing (e.g., conditionals). CT practices refer to the activities students engage in, like algorithmic thinking or testing and debugging, when creating computational projects. Finally, students need to develop empowered perspectives to feel confident that they can solve challenges in the world around them using computing and be active users of digital technologies.

More specificity in definitions of applied reasoning skills in CT has come from research and standards. For example, Grover and Pea (2013) provide details about the range of applied reasoning processes involved, including making abstractions and pattern generalizations; systematic processing of information; recognizing symbol systems and representations; implementing algorithmic notions of flow of control; engaging in structured problem decomposition (modularizing); employing iterative, recursive, and parallel thinking; applying conditional logic; observing efficiency and performance constraints; and engaging in practices of debugging and systematic error detection. As can be seen, the range of applied reasoning processes involved in CT is broad and complex, posing educators with a challenge of where to start and how to develop competence over time. The K–12 Computer Science Framework (2016), led by the Association for Computing Machinery, Code.org, and others, also emphasizes the thought processes in CT, identifying four specific CT practices, including recognizing and defining computational problems, developing and using abstractions, creating computational artifacts, and testing and refining computational artifacts.

In common to many CT definitions and frameworks is the concept of *practices*, or the *application* of design and inquiry to solving computational problems and creating computational artifacts. This reflects an orientation toward not just an internal, individual "thinking" but "ways of being and doing" that students should demonstrate when learning and exhibiting computer science knowledge, skills, and attitudes. Based on Jonassen's (2000) approach, one can also discern from these CT practices the rough outlines of the sequence of tasks: framing a problem in a way that a computer can solve it, defining the relevant algorithms and subroutines for implementing a solution to that problem, and refining the solution through iteration and debugging.

The centrality of the concept of practices to CT presents challenges to both computer science (CS) teachers in their instructional and assessment activities, and to ITS developers, who have to make complex decisions about how to model these practices in interactive assessment and tutoring systems. Secondary CS teachers, who are often underprepared to teach the subject, have limited curricular guidance on how to teach CS concepts and inquiry skills in a way that impacts student CT practices. More significantly, teachers are underprepared to effectively assess students' CT practices. One of the main challenges they face is modeling the relationships between what they want students to know, what counts as observable evidence of these skills, and how to develop tasks to elicit the evidence needed for drawing valid inferences about student performance. This same assessment modeling challenge is extended to ITS developers, who face a myriad of decisions about which components of CT practices to prioritize, what tasks to emphasize to teach the CS knowledge and skills, and how to embed interactive assessment and tutoring around learning CT reasoning. Principled assessment modeling processes and CT design templates can be helpful to support such work.

Science Inquiry: Importance and Construct Definition

In a similar fashion, science educators have developed a greater emphasis on inquiry skills over the past decade. The reasons for this shift are based in the increasing evidence that students learn science concepts best through applied reasoning, as exemplified in the quote: "Engaging in the practice of science helps students understand how scientific knowledge develops" (National Research Council, 2012, pg. 42).

This shift to inquiry has culminated in recent years with the publication of the Next Generation Science Standards (NGSS Lead States, 2013) by a consortium of states and the National Science Teachers Association, the American Association for the Advancement of Science, the National Research Council, and Achieve, a non-profit organization. The Standards offer a vision of what K–12 students need to know and be able to do to be scientifically literate and effective members of the US workforce. As of February 2016, 18 states and the District of Columbia had adopted NGSS, and many individual school districts have decided to introduce NGSS ahead of statewide adoption. The NGSS was designed with the goal of reflecting scientific activities within science education, consistent with calls from science educators and researchers over the past several decades (National Research Council, 2012).

To achieve science education that extends beyond the view of science as the accumulation of facts, the vision offered by the Standards weaves three dimensions: *scientific practices*, *disciplinary core ideas*, and *crosscutting concepts*. For the purposes of this chapter, it is important to note that the standards integrate applied science practices as a way for students to learn and a way for teachers to assess the development of students' understanding of core disciplinary ideas over time and students' capacity to discern crosscutting concepts.

The dynamic and applied quality of these standards is notable and worth reviewing in more detail. First, although the NGSS preserves the traditional disciplinary core ideas within the physical sciences; life science, earth and space science; as well as engineering, technology, and applications of science, it takes a purposely developmental view. Indeed, the developers of the Standards state that educator's role is "to prepare students with sufficient core knowledge so that they can later acquire additional information on their own " (National Research Council, 2012, pg. 31).

Second, this developmental perspective moves to another level in the crosscutting concepts, which "bridge disciplinary boundaries, having explanatory value throughout much of science and engineering" (National Research Council, 2012, p. 83). They are intended to provide students with a way to organize their understanding of scientific concepts and to see the connections across seemingly disparate concepts. Among the seven crosscutting concepts are cause and effect, energy and matter, and structure and function. For example, *structure and function* is defined as "the way in which an object or living thing is shaped and its substructure determine many of items properties and functions" (National Research Council, 2012, p. 84).

At the heart of the NGSS, however, are the scientific practices that are comprised of eight scientific and engineering inquiry processes: asking questions (for science) and defining problems (for engineering); developing and using models, planning and carrying out investigations, analyzing and interpreting data, using mathematics and CT, constructing explanations (for science) and designing solutions (for engineering), engaging in argument from evidence, and obtaining, evaluating, and communicating information. As can be seen when reflecting back on Jonassen's (2000) approach, one can again see that these practices unfold in a sequence that begins with framing a question or problem aligned with scientific principles, proceeds to defining the relevant models, investigative methods, and design approaches for addressing that question or problem, and concludes with engaging in a review of the evidence and developing the argument to communicate any results and findings.

The sheer complexity of these standards poses a formidable challenge for instruction and assessment that affects not only instructors, but also ITS developers. First, what are the best ways of integrating the three dimensions in tasks, such that every learning activity reasonably incorporates a core idea, practice, and crosscutting concept? Second, what science teaching and assessments can support student learning of this complex collection of knowledge and skill? One challenge of performance assessment is that they take considerable time for students to complete, with relatively lean information contained in students' responses. This "data poor" quality of performance assessment is in contrast with multiple-choice, which is a relatively efficient way of obtaining information on candidate performance (see Shavelson, Baxter & Pine, 1991, for a discussion of this issue). Thus, although technology enables a greater range of performance assessments than in the past, there remains the challenge of interpretation: how to make sense of the rich array of data that performance assessments make available about student solution processes (cf. Katz & Gorin, 2016; Keehner, Feng, Gorin & Katz, 2016). To obtain richer and denser information on student performances from a performance assessment, some assessment designers have turned to using process data – log files, telemetry, clickstreams, and other reflections of the moment-by-moment actions of students – to provide insights into students' knowledge and skills. However, interpretation of process data requires a deep understanding of student performance and the cognitive processes (i.e., knowledge and skills) that lead to that performance.

The challenge for ITS developers is, therefore, in moving from complex, interactive performances (i.e., observable behavior, or evidence) to drawing conclusions about what students know and can do (i.e., unobservable competencies), and to do so in a way that is consistent with prior results in the literature on cognition and learning. A principled and systematic assessment design process can help draw these connections.

Evidence-Centered Design (ECD)

As the previous section makes clear, the process of learning applied reasoning skills involves not only knowledge recall, but also the development of other skills, such as understanding the timing and situational constraints that support judgments about when applying knowledge is useful and appropriate. Its development is not linear, but more organic and dependent on opportunities for application. These aspects of applied reasoning pose a tremendous challenge to education and training, not so much from the perspective of how to sequence learning activities, but from the perspective of how to accurately detect where a given individual learner is in the process of learning to engage in applied reasoning. Seeing where the learner stands involves more than simply collecting a set of knowledge propositions and making sure the learner can articulate them; it involves having a way to see the full range of knowledge propositions and judgments that a learner generates over the course of changing situational constraints in any process of production or performance over time.

Fortunately, the ECD model specification approach is well suited to such modeling challenges. At the high level, it organizes the entire process of documenting complex reasoning as an assessment argument. This argument has three parts: the student model, the task model, and the evidence model. Despite the similar terminology to ITS design, the actual operational meanings of these three models are distinct.

While in ITS development, the term "student model" represents the inferred "state" of the learner's knowledge and skill at any given time in the learning process, in ECD modeling, student model generally refers what knowledge and skills are being taught or measured. This student model can reflect 1) the desired end state of the learner's knowledge and skills in a domain (e.g., the expert's mind) and/or 2) a series of transitional developmental states that the learner takes on the path from novice to expert. It is derived from the processes of domain analysis and domain modeling (described later). Another core element of ECD is the evidence model, which describes how to measure the desired knowledge and skills in terms of learners'

observable performances. Finally, the task model describes the attributes of tasks that will elicit the observable evidence of measured knowledge and skills. The specific details of student, evidence, and task model are initially recorded in documents called design patterns, which may be adapted to apply across different domains. ECD helps designers link features of tasks to particular performances and how they yield evidence that specific types of knowledge and skill have been learned (Mislevy & Haertel, 2006).

In practice, the ECD model specification process is typically described in terms of five layers: 1) domain analysis, 2) domain modeling, 3) conceptual framework, 4) implementation, and 5) delivery (Figure 1).

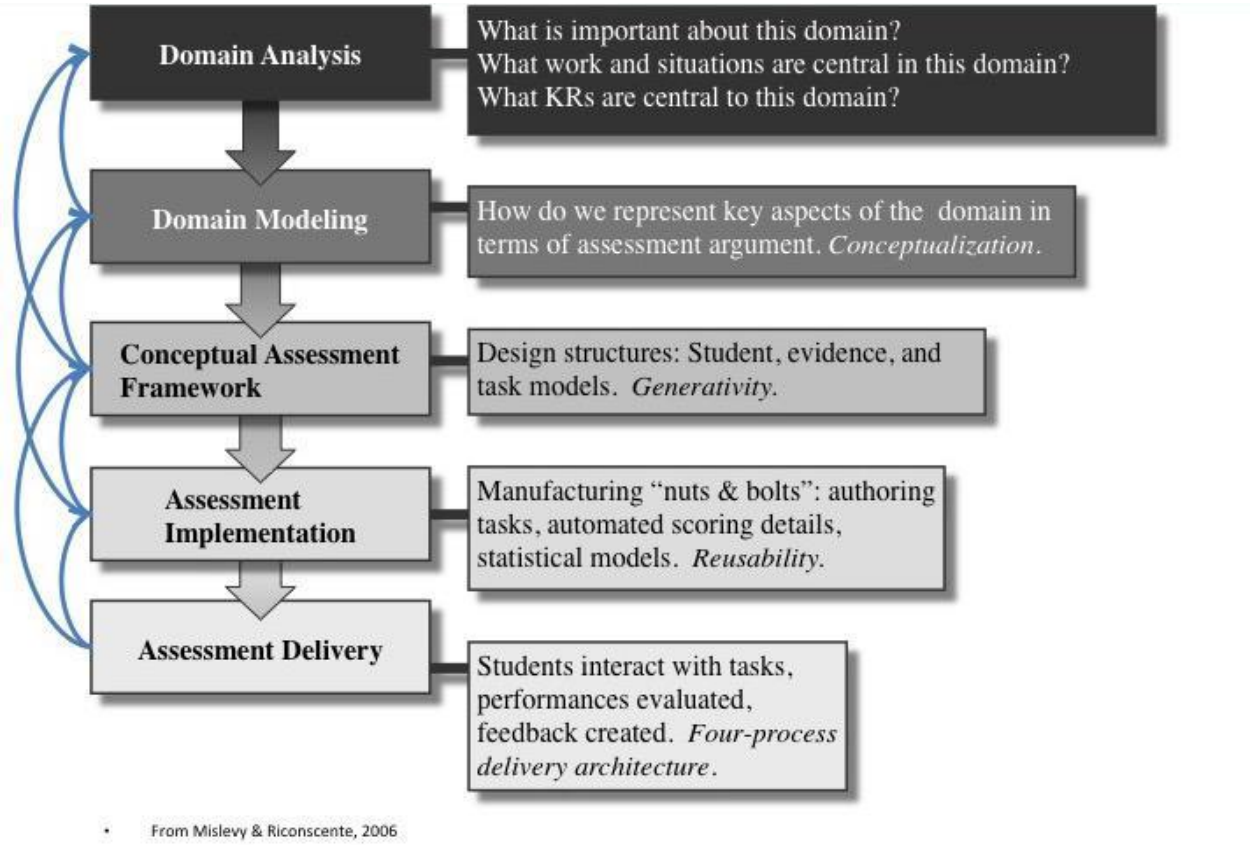


Figure 1. Layers of the ECD process

Although the layers suggest a sequential design process, cycles of iteration and refinement are intended, both within and across layers, and work in different layers can be pursued simultaneously. The focus of this chapter is on the first two layers of ECD, which represent the critical methods of documenting the knowledge, skills, and abilities in complex forms of reasoning. The additional layers build on these first two, and focus more on the development and implementation.

Domain Analysis

In domain analysis, developers and experts identify and analyze the domains, constructs, and underlying skills of interest, as well as relevant standards and benchmarks (if available). The analyses examine the ways people acquire and use the knowledge and skills, situations under which knowledge is used, and

indicators of successful application of the knowledge. Typical sources of information include existing domain and construct definitions, curriculum documents, and relevant research findings (such as those that support a validity argument or provide examples of assessment tasks). Typical outcomes of the domain analysis include lists of organizing concepts and principles in the domain(s).

Domain Modeling

Through domain modeling, developers organize information from the domain analysis as a design pattern (Mislevy, Steinberg & Almond, 2003). Design patterns are structured narratives explicating an assessment argument. Assessment arguments specify the knowledge, skills, and abilities (KSAs) one wants to address, potential observations and work products that can provide evidence about acquisition of this knowledge or skill, potential rubrics to evaluate student performances on the tasks, and features of task that enable the student to provide this evidence (Messick, 1989; Mislevy and Haertel, 2006). When completed, a design pattern makes the relationships among these attributes explicit.

Modeling CT Practices

To illustrate how ECD has been used to model CT practices (CTPs), this section presents an example of a design pattern that was developed by SRI International as part of a research project that designed assessments for CT in K12 education, specifically early high school: Principled Assessment of Computational Thinking (PACT).

The example design pattern was developed as part of a larger set of design patterns created to support assessment of CTPs across different curricula for CT. The CTP design patterns were developed in 2011 through consultation with various experts and working groups in computer science education and assessment, and close examination of STEM construct definitions; standards; curriculum; literature in computer science education, scientific inquiry, engineering design, communication, and collaboration; and previous CS assessment projects (see Bienkowski, Snow, Rutstein & Grover, 2015). Special attention was paid to the Computer Science Teacher Association (CSTA) standards, the *Exploring Computer Science* (ECS) curriculum (including learning objectives), and the AP Computer Science Principles framework (including learning objectives and evidence statements).

Table 1 shows the core attributes of one of these ECD design patterns for CTP: design and apply abstractions and model. The design pattern attributes begin with an overview of the CT practice, which provides a narrative summary of the key features of the construct. The next attribute consists of the focal knowledge, skills, and attributes (FKSAs) associated with the CTP and form the target outcomes for assessment. In this case, the design pattern lists a subset of KSAs that are associated with what is considered foundational knowledge involving abstraction in computing. These FKSAs, as well as others associated other construct components (analyzing a model or abstraction), form the core documentation material for the *student model* of the assessment. The attributes continue with the potential observations, which comprise the observable behaviors showing evidence of competence, and potential work products, which describe the kinds of tasks that elicit observable evidence of competence in the target outcomes (FKSAs). This is the core documentation material for the *evidence and task models* of the assessment, and it is listed here because understanding what one hopes to see will help frame the design of the task and rubrics in the subsequent sections of the design pattern.

Table 1. FKSAs, potential observations, potential work products. CTP design pattern: design and apply abstractions and models.

Overview
Thinking strategically about abstraction is a hallmark of computational thinking. This design pattern supports the development of tasks in which students use ideas and representations that capture general to specific aspects, or patterns, of an entity or a process and the relationships/structures among entities or processes, including level of detail. This may include designing general solutions to problems or generalizing a specific solution to encompass a broader class of problems (functional abstraction). These ideas and representations may be used in different contexts (problem or disciplines). Students demonstrate knowledge of the representational properties of discrete mathematics, models, diagrams, computer programs (data abstraction), items found in the natural and manmade world, and others. They also demonstrate an understanding of the limitations of models to represent phenomena and attention to the purpose of the model or abstraction.

FKSAs	**Potential Observations**	**Potential Work Products**
• Ability to explain what abstraction is, both functional and data. • Ability to reason about a problem at multiple levels of detail. • **Ability to explain the benefits of using abstraction in problem solving, e.g., to manage complexity and generalize patterns.** • Ability to explain that an algorithm is a form of abstraction that contains a sequence of instructions whose end state or output can be determined once given a particular starting state. • Ability to explain the characteristics of problems for which abstraction would be useful. • Ability to describe how a computer model makes a representation of the real world. • Ability to explain how computers represent mathematical objects and logical operations for purposes of computation and modeling. • Ability to explain how computers represent objects as data and data as objects (e.g., media files, QR codes). • Ability to explain the connections between elements of mathematics and computer science including binary numbers, logic, sets, and functions.	• **The degree to which the abstraction matches the need of the problem** • The accuracy of the representation • The number of representations used • The appropriateness of the representations • The appropriateness of the explanation • The degree to which the implementation matches the abstraction • The degree to which abstraction is applied in the implementation • The degree to which the map is appropriate for the problem and the elements • The degree to which the explanation or documentation is a clear description of the abstraction (clarity of the explanation or documentation) • The degree to which the analysis is appropriate • The correctness of the analysis • The appropriateness of the application or the correctness of the application of the abstraction • The accuracy of the implementation or application of the abstraction	• One or more representations of an abstraction, problem, problem space, or analysis • **The explanation of or related to the abstraction (such as how it was applied, why it is appropriate)** • Implementation of the abstraction • A mapping between elements of the problem and elements of the abstraction • Explanation or documentation of the implementation or abstraction • The analysis of the abstraction or model • The application of the abstraction

Design patterns typically include several additional attributes, including characteristic task features, rubrics, and variable task features, attributes designed to help the designer further refine the assessment tasks, including essential features of all tasks measuring the practice, and features of the task that can be modified to rebalance FKSA coverage or the difficulty of the task. Note that all attributes in a design pattern are specified in narrative form, which makes them more accessible to cross-disciplinary assessment development teams.

One of the primary benefits of using design patterns comes from their relational structure. This relation is illustrated in the boldface statements in the design pattern depicted in Table 1. In this simple example, if one wants know whether a student can explain how abstraction and modeling can help in computational problem solving then they need to develop an assessment task that elicits an explanation of how an abstraction is appropriate for solving a specific problem, or group of problems. Importantly, it also specifies the general type of evidence that is appropriate for making inferences about how well a student can explain how abstraction and modeling can help in computational problem solving, in this case, the degree to which the selected abstraction (in the explanation) matches the stated problem (e.g., why it is appropriate). The specific types of evidence will be further defined in one or more rubrics that will need to be applied to score the explanations.

Using this design pattern, an assessment designer can insert questions into learning activities that require the development, justification, and explanation of abstractions of inherent processes. For example, learners may be asked to create a program that will solve a problem—and this creative process may unfold over several phases of the extended activity. For example, the learner may be asked to engage first in a task that requires identifying a real-world problem that a computer might be able to solve. Later in the design process, the learner may be expected to implement an algorithm (abstraction) in a computer programming language or environment that will produce the product or solution. In this stage, the example design pattern above may be most important to use for assessments.

As this example illustrates, the narrative and relational structure of the example design pattern permits instruction and assessment designers to document and relate both the cognitive and metacognitive components of CTPs for the purposes of assessment design. Such an approach permits a designer to clearly relate evidentiary performance data and ways of eliciting that data with each distinct component of CT.

Modeling Science Inquiry Practices

As described in the introduction of this chapter, the NGSS document (NGSS Lead States, 2013) frames science instruction as incorporating three intertwined dimensions: science practices, disciplinary core ideas, and crosscutting concepts. Creating appropriate instruction and assessments aligned with NGSS requires further detail. Researchers at the Educational Testing Service (ETS) have used ECD modeling over the past several years to develop a large-scale science competency model and associated evidence models and tasks that help developers create NGSS assessments (Liu, Rogat & Bertling, 2013).

The researchers conceptualized the ECD competency and evidence models to focus not just on one level, but multiple levels of science achievement. This NGSS collection of gradually more sophisticated understandings is called a "learning progression." The resulting indicators focus on evidence – "what students know and can do" – instead of general standards – "what students *should* know."

For example, target knowledge at the lowest level (level one) of the learning progression model might be defined as reflecting a very basic understanding of key concepts and may demonstrate particular misconceptions. Knowledge at an intermediate level (level three) of the learning progression model might include more sophisticated skills and concepts, and knowledge at the highest level would reflect complete and correct knowledge and skills. At each level, the learning progression might also contain instructional strategies for helping students to move to the next level. Thus, a learning progression is a type of ECD model that defines attributes of both the student model and the evidence model, containing both descriptions of student competencies (whether knowledge, skills, or both) as well as statements of what students can do (i.e., evidence of those competencies).

Table 2 shows two levels of a learning progression (LP) for "matter and its interactions" (Liu et al., 2013), one of the NGSS disciplinary core ideas, and Table 3 shows an LP constructing explanations from evidence, for one of the science practices.

Table 2. LP document for understanding of properties of matter.

Achievements	Gap/Challenge	Instructional experience to support progression
Level 1—Macroscopic compositional model (Not shown)		
Level 2—Microscopic compositional model (Not shown)		
Level 3—Developing particle model		
• Conceives of matter as made of particles that have mass and volume. • Sometimes thinks of empty space between particles for some materials. • Sometimes recognizes that particles of matter move.	• May not recognize that there is empty space between particles in all conditions, although having a nanoscopic notion of material identity. • May not recognize particles move for all substances and all states of matter.	• Help students understand that different properties of matter are determined by the arrangement and motion of particles making up the matter by using computer simulations. Provide investigative opportunities for students to explore the relations between properties of matter and the arrangement and/or motion of particles.
Level 4—Particle model		
• Consistently conceives of matter as made of particles that have mass and volume. • Consistently thinks there is empty space between the particles. • Consistently recognizes that particles of matter move. (Note: students are not expected to think about absolute zero.) • Recognizes that temperature is a product of the average kinetic energies of the particles of the substance.	• May not consistently recognize that different materials are made of specific atoms, or combinations of atoms forming molecules, although may have a general particle model.	• Provide investigative opportunities for students to construct arguments about the behaviors of matter undergoing chemical change by using an atomic-molecular model to evaluate which argument better explains and predicts chemical change.
Level 5—Atom-molecular model (Not shown)		

Table 3. LP document for constructing explanations.

Achievements	Gap/Challenge	Instructional experience to support progression
Level 1—Nonstructural mode (Not shown)		
Level 2—Noncausal relation model (Not shown)		
Level 3—Insufficient causal relation model		
• Student makes an accurate claim. • Student backs up the claim with evidence.	• The evidence is insufficient / inappropriate	• Make explicit to students what counts as appropriate and sufficient evidence (i.e., by appropriate, we mean data that are relevant to the problem and help support the claim; sufficient refers to providing enough data to convince another individual of the claim). Often providing sufficient evidence requires using multiple pieces of data. • Model the justification of claims with sufficient evidence with examples. • Draw on what students know about evidence or justification in their everyday life and help them understand what counts as good evidence. • Provide feedback in response to students' justifications.
Level 4—Causal relation model		
• Student makes an accurate claim. • Student backs up the claim with sufficient and appropriate evidence.	• Student does not use reasoning to tie the claim and evidence together.	• Help students understand why it is important to include reasoning to convince others to accept the claims. • Make explicit to students what reasoning is (i.e., reasoning links the claim and evidence and shows why the data count as evidence to support the claim). • Model the use of reasoning with examples that tie claims to evidence. • Draw on what students know about reasoning in their everyday life and help them understand scientific reasoning. • Provide feedback in response to students' reasoning.
Level 5—Insufficient coherence model (Not shown)		
Level 6—Sufficient coherence model (Not shown)		

In applied reasoning tasks and performances in science inquiry, learners often are engaged in multiple facets of knowledge and skill defined by these multiple ECD models. For example, they might be asked to explain how heat affects water during a classroom experiment, and their explanation may be judged according to the criteria of accurate conceptual understanding of the properties of matter (as partially illustrated in the LP shown in Tables 2 and 3) and according to criteria of well-supported explanations as defined in the LP for scientific explanations. The LP model has the added benefit of coordinating how learners may progress differently on conceptual understanding and specific practices.

Discussion

GIFT researchers emphasize the benefits of standardized approaches to the development of adaptive or automated instructional materials. This chapter has provided some illustrations of how ECD modeling can support such standardized approaches in the design of assessments of the most complex forms of applied reasoning and problem solving. The benefits of ECD modeling focus on its theoretical soundness, consistency of documentation format, and potential for model adaptation across multiple domains.

The design patterns and learning progression models presented in this chapter demonstrate the potential for developing a GIFT library of competency models that are thorough, detailed, and grounded in empirical learning science. The ECD documents have common attributes based on the highest standards of measurement theory and psychometrics, and as such, provide a common language and framework that may be used by subject-matter experts and ITS developers in multiple domains.

Further, the ECD documents are designed to address as many different components of knowledge and skill as is needed to accurately capture the complex processes of applied reasoning and problem solving in any given domain. This is possible because ECD documents lend themselves to being combined in flexible ways to address the unique conditions of applied reasoning in different domains. We have highlighted just two combinatorial schemes in this chapter: combinations that support *phased assessment* of applied reasoning as it unfolds in over time (the CT case), and combinations that support *differentiated assessment* of distinct knowledge and skill progressions involved in applied reasoning (the science inquiry case).

Recommendations and Future Research

Future research should examine what elements of the ECD modeling approach can be integrated into the GIFT platform and its standardized templates. Researchers face several challenges to developing efficient methods of defining and recording the attributes for automated assessment of complex reasoning and problem solving in ITSs. While ECD has laid the foundation for accurately recording the prevailing consensus of subject matter experts on core KSAs and their development, the narrative quality of these recording methods requires ITS designers to make many additional decisions and trade-offs, many of which continue to be made based on convenience rather than optimization. Too little is known about how to prioritize, integrate, and balance aspects of assessment across complex tasks to support improved competence. Future research needs to focus on 1) refining methods for representing the core KSAs in a manner that permits easy interpretation and use by designers of automated assessment systems; 2) defining a set of meta-design principles needed to support assessment selection and integration across complex tasks, including consideration of the appropriate balance between covert and overt formative assessment; and 3) finding ways to provide guidance to ITS designers about the optimal levels of assessment feedback and monitoring to support learner progress in applied reasoning over time.

To maximize the potential for standardized design of a modular specification template for applied reasoning and problem solving, such research might occur concurrently across one or more domains and/or component applied reasoning skills.

Acknowledgements

We would like to dedicate this chapter to the memory of their late SRI International friend and colleague, Geneva Haertel. Dr. Haertel was a distinguished assessment researcher, generous mentor, and longtime ETS collaborator. She passed away in late summer 2016 and was to have co-authored this chapter. Her scholarship was foundational to the ideas we present and her intellectual spirit guided us throughout its writing.

Some of the work reported in this chapter is based on SRI International's Principled Assessment of Computational Thinking (PACT) project, which as conducted by the Center for Technology in Learning at SRI International with support from the National Science Foundation (NSF) under contract numbers, CNS-1132232, CNS-1240625, DRL-1418149, and CNS-1433065. Any opinions, findings, conclusions, or recommendations expressed in this chapter are those of the authors and do not necessarily reflect the views of the NSF.

References

Adams, J. B. (2008). Computational science as a twenty-first century discipline in the liberal arts. *Journal of Computing Sciences in Colleges, 23*(5), 15–23.

Astrachan, O., Hambrusch, S., Peckham, J. & Settle, A. (2009). *The present and future of computational thinking*. Paper presented at the Proceedings of the 40th ACM Technical Symposium on Computer Science Education, Chattanooga, TN. doi: 10.1145/1508865.1509053.

Bienkowski, M., Snow, E., Rutstein, D. W. & Grover, S. (2015). *Assessment design patterns for computational thinking practices in secondary computer science: A first look* (SRI technical report). Menlo Park, CA: SRI International. Retrieved from http://pact.sri.com/resources.html.

Brennan, K. & Resnick, M. (2012). *New frameworks for studying and assessing the development of computational thinking*. Paper presented at the 2012 annual meeting of the American Educational Research Association (AERA), Vancouver, B.C.. Retrieved from: http://web.media.mit.edu/~kbrennan/files/Brennan_Resnick_AERA2012_CT.pdf.

Butler, D. L. & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. doi:10.3102/00346543065003245.

Denning, P. J. (2009). The profession of IT Beyond computational thinking. *Communications of the ACM, 52*(6), 28–30.

Desmarais, M. C. & Baker, R. S. J. D. (2012b). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*(1), 9–38. http://doi.org/10.1007/s11257-011-9106-8.

Galyardt, A. (2015). Move your lamp post: Recent data reflects learner knowledge better than older data. *Journal of Educational Data Mining, 7*(2), 85–111. Retrieved from http://educationaldatamining.org/JEDM13/index.php/JEDM/article/view/JEDM100.

Grover, S. & Pea, R. (2013). Computational thinking in K–12: A review of the state of the field. *Educational Researcher, 42*(1), 38–43. Retrieved from: http://people.cs.vt.edu/~kafura/CS6604/Papers/CT-K12-Review-State-Of-Field.pdf.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational technology research and development, 48*(4), 63–85.

Kafai, Y. B., Peppler, K. A. & Chapman, R. N. (2009). The Computer Clubhouse: Constructionism and Creativity in Youth Communities. *Technology, Education--Connections*. Teachers College Press. 1234 Amsterdam Avenue, New York, NY 10027.

K–12 Computer Science Framework Steering Committee. (2016). *K–12 computer science framework*. Retrieved from http://www.k12cs.org.

Katz, I. R. & Gorin, J. S. (2016). Computerising assessment: Impacts on education stakeholders. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 472–489). New York: Routledge.

Keehner, M., Gorin, J. S., Feng, G. & Katz, I. R. (2016). Developing and Validating Cognitive Models in Assessment. In A.P. Rupp & J.P. Leighton (Eds.), *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (pp. 75–101). New York: Wiley-Blackwell.

Koedinger, K. R., Brunskill, E., Baker, R. S. J. d., McLaughlin, E. A. & Stamper, J. (2013). New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine, 34*(3), 27–41. doi: http://dx.doi.org/10.1609/aimag.v34i3.2484

Liu, L., Rogat, A. & Bertling, M. (2013). *A CBAL(TM) science model of cognition: Developing a competency model and learning progressions to support assessment development* (ETS Research Report No. RR-13-29). Princeton, NJ: Educational Testing Service.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5–11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher, 23*(2), 113–23.

Mislevy, R. J. & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20.

Mislevy, R. J. & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61–90). Mahwah, NJ: Erlbaum.

Mislevy, R.J., Steinberg, L.S. & Almond, R.G. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development* (pp. 97-128). Mahwah, NJ: Erlbaum.

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: National Academies Press.

National Research Council (2010). *Report of a workshop on the scope and nature of computational thinking*. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/.

National Research Council. (2012). *Report of a workshop on the pedagogical aspects of computational thinking*. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/.

Park, O. & Lee, J. (2003). Adaptive instructional systems. In D. H. Jonassen & M. Driscoll (Eds.), *Handbook of research for educational communications and technology* (2nd ed., pp. 651–684). Mahwah, New Jersey: Laurence.

Pintrich, P. (2000). An achievement goal theory perspective on issues in motivation terminology, theory, and research. *Contemporary Educational Psychology, 25*(1), 92–104. doi:10.1006/ceps.1999.1017

Shavelson, R. J., Baxter, G. P. & Pine, J. (1991). Performance assessment in science. *Applied measurement in education, 4*(4), 347–362.

Smith, M. (2016). *Computer Science For All*. Washington, DC: The White House. Retrieved December 21, 2016, from https://www.whitehouse.gov/blog/2016/01/30/computer-science-all.

Sottilare, R. A., Graesser, A., Hu, X. & Holden, H. (Eds.). (2013). *Design Recommendations for Intelligent Tutoring Systems: Volume 1-Learner Modeling* (Vol. 1). US Army Research Laboratory.

United States Department of Education. (2015). *Fact sheet: Testing action plan*. Washington, DC: US Department of Education. http://www.ed.gov/news/press-releases/fact-sheet-testing-action-plan.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33–35.

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical transactions of the royal society of London A: mathematical, physical and engineering sciences, 366*(1881), 3717–3725.

Winne, P. H. & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. Pintrich & M. Zeidner (Eds.) *Handbook of self-regulation*. (pp. 531–566). San Diego, CA, US: Academic Press. http://dx.doi.org/10.1016/ B978-012109890-2/50045-7.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B.J. Zimmerman & D. H. Schunk, (Eds.) *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 145–178). Mahwah, NJ: Erlbaum. Retrieved from http://tinyurl.com/Hkewfmb.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary educational psychology*, *25*(1), 82–91.

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 1–37). Mahwah, NJ: Erlbaum. Retrieved from http://tinyurl.com/hkewfmb.

CHAPTER 8 – Assessing Individual Learner Performance in MOOCs

Ryan S. Baker[1], Piotr Mitros[2], Benjamin Goldberg[3], and Robert A. Sottilare[3]
University of Pennsylvania[1], edX[2], US Army Research Laboratory[3]

Introduction

Massive open online courses (MOOCs) have emerged as a prominent mode of online education, but the quality of assessments in MOOCs remains inconsistent. There has been a consistent gap between the state of the art in assessment and the state of the practice in both MOOCs and other forms of educational technology. Improving the quality of assessment has the potential to improve their usefulness for certification, formative feedback, and learning. In this chapter, we discuss this gap and efforts to improve assessment in MOOCs. There are several potential directions for improving assessments in MOOCs, including improving the psychometric properties of simple assessment types, such as multiple-choice and fill-in-the-blank questions; creating richer assessment experience through more student interaction and more engaging experiences (e.g., games, and simulations), through leveraging help resources to see how students perform with some degree of support; and using more powerful technology such as automated essay scoring to better assess students. We discuss both existing efforts and ways that research in other communities can be incorporated into MOOC platforms.

In fall 2011, a small group of computer scientists from Stanford University launched the first xMOOC (an experimental MOOC based on a traditional university course), the Stanford AI Course. This early MOOC platform only supported three assessment types – multiple choice questions, select-many multiple choice questions, and numeric answers. However, the course used these types of assessment in a relatively sophisticated way. Some assessments were embedded in the video and tightly integrated with content presentation, giving continuous formative assessment throughout the course. As instructors walked through mathematical derivations, students would do portions of the work (this practice is sometimes referred to as a partially faded worked example – e.g., Salden et al., 2010). In other assessments, students were given a problem, asked to develop an algorithm to solve that problem, and asked to enter the output of their program implementing that algorithm into a numeric response question. This resulted in a relatively rich student experience.

Shortly after the Stanford course, Coursera introduced two courses, one in machine learning and the second in databases. Coursera's initial courses relied more on simple multiple choice questions, although they introduced problem banks to create the possibility of mastery learning through selecting items from the problem banks and continuing to offer items until the student demonstrated mastery. In addition, Coursera introduced assignments that could be graded via external software, a technology mostly used for programming assignments.

The MITx platform, now Open edX, was introduced approximately 6 months later. In the first MITx course, students worked through design and analysis problems which they could answer with numbers, equations, and circuit schematics. Their answers were verified with circuit simulations (in a JavaScript Spice clone), Python programs, plugging numbers into equations, or comparison of numbers with tolerance. Even with numeric answers, there were often multiple correct answers. For example, a design problem that required students to pick component values for a resistive divider of a given attenuation would need to verify 1) ratios of those components, 2) whether they were valid values, and 3) whether they were realistic values (Mitros et al., 2013). This type of nuanced complexity to student solutions is not unique to this course, but is found in a range of higher education course assessments, particularly in domains such as engineering.

The first dozen edX courses launched with several assessments that were discipline-specific, such as tools for evaluating computer code, chemical equations, and visualizing crystallography planes.

Around the same time, Coursera introduced calibrated peer assessments where an essay or other artifact is graded by a student's peers, allowing for human assessment at a scale that would not be feasible for an instructor or teaching assistant in a course with tens of thousands of students.

However, despite the availability of support for programmatic auto-grading, calibrated peer assessment, and other forms of rich assessment embedded into teaching activities, most MOOCs rely upon relatively simple forms of assessment. The majority of MOOCs rely upon weekly quizzes or assignments where the student answers a fixed set of multiple-choice or fill-in-the-blank questions. While edX has support for over 50 different types of activities available and ready to use as options for course authors, around 80% of assessments are either multiple choice questions or similar simple assessments. Most competing platforms rely even more heavily on multiple-choice questions.

When compared to the rich and broad range of assessment seen within computer-based learning environments, for example, especially within the research community, the limitations of current practice in MOOCs and many other digital learning technologies, become especially notable. While learners using computer-based learning environments such as Cognitive Tutor are assessed at the same time they are learning (Corbett & Anderson, 1995), assessment in MOOCs is often separate from learning. For example, although edX is designed around interleaved learning sequences with integrated formative assessment, only half of the learning sequences with videos in courses deployed on edX.org include integrated assessments. While adaptive assessments in systems like the Assessment and Learning in Knowledge Spaces (ALEKS) determine where a student is in the progression from learning prerequisites to learning the skills that build on those prerequisites (doignon & Falmagne, 2012), MOOC developers rarely have the resources to create sufficient numbers of assessments to enable adaptive pathways, especially given the thousands of courses used at a university level. Consequently, adaptivity is limited to feedback and hinting. While learners using simulations take part in rich performance assessments (Clarke-Midura et al., 2012), MOOC learners often are assessed solely on multiple choice and fill-in-the-blank items, since such assessments have not yet been developed for the diversity of topics covered in MOOCs. Many examples of richer assessment can be found. Applying these approaches to the scale and diversity of courses offered as MOOCs remains an open challenge. The problem is complicated by the types of instructors teaching MOOCs – in contrast to trained educators at a K–12 level, university courses (and the MOOCs based on them) are generally taught by subject-matter experts, with little background in teaching and learning or educational technology.

In this chapter, we discuss the efforts that have been made to enhance assessment within MOOCs and potentials for enhancing assessment in MOOCs further. As MOOCs continue to serve tens of millions of students with thousands of courses from hundreds of universities, enhancing assessment in this context has the potential to improve learning experiences for many people in many contexts.

Improving the Quality of Current Common Types of Assessments

As mentioned previously, many MOOCs assess learners with a fixed set of multiple-choice and fill-in-the-blank items. Even within the paradigm of this type of item, there are several limitations to current practice in this regard. One limitation is that items are typically not validated. Item validity in assessment construction is a well-known area, incorporating both the psychometric properties of items (Dudycha & Carpenter, 1973) and their mapping to the content goals for the assessment (Mislevy & Riconscente, 2006). However, applying such algorithms in MOOC settings is still an area of open research (Champaign et al., 2014). There are several challenges to integrating such algorithms into MOOCs:

- When courses rely on mastery learning, students have much more incentive to guess and explore. Incorrect answers are often a result of such exploration, rather than item difficulty, as assumed in models such as the Rasch model or other Item Response Theory (IRT) models.

- MOOCs are open, and many learners have no intention of completion. Many students only target sections of courses that are relevant to them. More advanced learners are likely to skip over or rush through easy items, while novices might skip difficult problems, which can reduce the accuracy of knowledge estimation.

- In contrast to an exam where knowledge stays constant during the course of the exam, or adaptive systems that alternate between learning and assessment (such as ALEKS), students in MOOCs learn over time, and items often have intellectual cohesion and sequential narrative. Models which presume learner knowledge is fixed, such as IRT, fail to work in such cases. Another popular knowledge modeling algorithm, Bayesian knowledge tracing (BKT), assumes that learners learn, but assumes a certain level of independence in the order of items. Learners self-regulate when working through MOOCs, invalidating such assumptions.

- MOOC assessments rarely have a 1-1 mapping to learning components, unlike psychometrically designed tests. Analysis and design courses will have complex, multi-concept questions, where a single problem might take several hours to complete (Seaton et al., 2014), while history courses might spot-check facts, such as asking about a key date. Many psychometric models are undercon-strained in such cases.

- MOOC offerings typically decline in size with each run, where the first run is substantially larger than successive ones. Applying data from the first run to improve future runs would not impact the large number of students in the first run, so improvements would either need to be developed based on a prototype cohort[1], or run in real time by analyzing and attempting to enhance problems after a minority of students attempted them (as found in systems such as the Learning Online Network with Computer-Assisted Personalized Approach [LON-CAPA]).

Another key difference is that MOOC assessments typically grade students with regard to percentage correct. Psychometricians have been aware for decades that items have different properties – even in the same topic area, items may have different difficulties and different degrees of discrimination (how effectively they distinguish skillful students from non-skillful students), and as such, grading according to percentage correct is far from optimal for measurement of student skill. However, psychometric considerations must be balanced with other goals of exercises in courses:

- Exercises serve as a primary means of learning. Mastery learning is a key technique for effective learning, repeatedly shown to lead to better learning outcomes (Bloom, 1987). In mastery learning settings, students are asked to continue to attempt problems until they demonstrate mastery, for example by getting a sufficient number of problems correct or according to the assessments of knowledge modeling algorithms (e.g., Anderson et al., 1995). This is, in many ways, opposite to the goals of measurement, where items students answer correctly 50% of the time contribute maximum information. A question that all students answer correctly is often an excellent tool for learning, but contributes little psychometric information.

[1]Many MOOCs first run in a residential format, and some MOOC providers have experimented with a prototype cohort with fewer students for validation of courses.

- Grading is used to motivate students. For those purposes, simplicity is paramount; students ought to understand grading schemes. Algorithms such as multivariate IRT are opaque to students.

- Grading is used for certification. In this context, it is important for users of such accreditation to understand what the accreditation means.

Nonetheless, MOOCs give a new opportunity for grading schemes which integrate psychometric considerations into learning engineering and learning design. Some pilot work has attempted to take these several issues into consideration within the context of MOOCs – for example, Colvin et al. (2014) uses a form of IRT to better assess students in a physics MOOC, taking item difficulty and discrimination into account. Pardos and colleagues (2013) similarly extend BKT to assess the degree of student learning over time in a MOOC on circuit design, taking item difficulty and multiple student attempts into account. However, these innovations have not scaled to the broader range of MOOCs or led to better achievement of the non-psychometric goals of exercises and assessments in MOOCs. How to do so effectively, what issues will come up with the diverse range of courses and pedagogies in MOOCs, and how to balance the competing design goals of assessments all remain open problems.

These problems are unlikely to be solved until tools that ingest MOOC data are standardized and widely deployed. Instructors and course development teams generally do not have the capacity to do this type of research and enhancement on their own. There are a number of initiatives to create such tools (Cobos et al., 2016; Dernoncourt, 2013; Fredericks, 2016), but none are standard or complete. A widely useful platform for open, integrated learning analytics (Siemens et al., 2011) remains a dream, for both technical and non-technical reasons. Scaling existing learning algorithms to perform the complex calculations needed for classical psychometric modeling on the terabytes of data in MOOCs in real time remains a technical challenge, while sharing and integration of proprietary learning data are open legal and policy questions.

Broadening Assessment in MOOCs

Another way that MOOC assessment could be enhanced is through broadening formats for assessment. As noted previously, the first xMOOC alternated between providing conceptual content in the form of video lectures, having students manipulate information, and assessing student understanding. Many xMOOCs have included rich analysis-and-design problems. While the functionality to do this remains present in the primary xMOOC platforms, it is underutilized; at least half of MOOCs do not make adequate use of such functionality[2].

A common strategy in K–12 intelligent tutoring systems (ITSs) is to build problem sets where a larger problem is broken down into different steps, and the student receives feedback at each step. This is broadly the strategy used in the highly successful mathematics ITS, Cognitive Tutor (Anderson et al., 1995). An example of these step-based intelligent tutoring problems are seen in Aleven et al. (2015), where a data science MOOC's assessments were converted into step-based problems. This allows more frequent feedback to students, and better tools for understanding student learning and knowledge gaps. However, it does not easily apply to all problems that might be seen in MOOCs. Many university courses strive for more complex, multi-concept authentic assessments, where students must not only work through calculations, but come up with a high-level problem-solving strategy in an open-ended setting. In an intermediate approach, students work in the platform in open-ended tools such as word processors, circuit schematic entry tools, or code editors, and the platform monitors student work. Traces of such activities mined for data.

[2]Approximately half of learning sequences on edX.org with videos do not include assessments. This number may be even lower on other MOOC platforms.

However, such systems are extremely expensive to build since such analyses tend to be very domain-specific. This engineering cost would be prohibitively expensive for the thousands of courses offered at a tertiary level.

Additional types of problems may afford richer experiences, both in terms of learning and assessment. For example, in the ITS Betty's Brain, students create concept maps to explain the interrelationships in a domain, and then the concept maps are evaluated in terms of their match to an expert-generated concept map (Leelawong & Biswas, 2008). Bringing such functionality into MOOCs is complex. Several MOOCs have encouraged learners to generate concept maps (e.g., Viswanathan, 2012; Bachelet et al., 2015), but MOOC developers have not yet leveraged the opportunity to automatically assess and provide immediate feedback on the resultant concept maps. There are several issues with bringing such tools to scale:

- Such tools tend to be complex to develop, both technically and pedagogically. University courses are taught by subject matter experts, often with little background in teaching-and-learning, educational technology, or computer programming. Although edX.org has over 50 activity types available to course authors, the Open edX ecosystem has at least twice that number, and many more are integrated through Learning Tools Interoperability (LTI), iframes, or JavaScript, only a minority of course teams are able to make use of such functionality.

- Developing such tools is expensive. In the tertiary space, there are thousands of courses[3], each taken by thousands or tens of thousands of students. Developing custom technology for each of these would cost hundreds of millions of dollars.

- Even seemingly broadly applicable tools, such as the concept maps mentioned, only apply to a minority of courses. Many university courses cover areas of active research, and as such have poorly-defined concepts and learning objectives. In these courses, the curriculum and objectives are still being defined.

A promising area of research is finding and encouraging the use of simple tools – both from a developer and course author standpoint – to enable cognitively complex tasks such as the concept maps seen in Betty's Brain. For example, in the context of learning-by-teaching, MOOCs have used Q&A forums, peer feedback tools, and community TAs. With thousands of mature students in a course, it is often possible to come up with relatively simple techniques which mirror the cognitive processes of students in ITSs, but do so by relying on either the intelligence of crowds or the intelligence of individual students.

Perhaps even richer interactivity and assessment can be found in systems that allow students to enter answers to conceptual questions in natural language, such as AutoTutor (Graesser et al., 2005). AutoTutor uses natural language processing (NLP) both to evaluate student responses and to ask probing questions that help to explore how much students understand. This type of assessment, while expensive to create, can help to richly explore student understanding, toward offering more sensitive responses and support to students. However, it is relevant to ask how to build out this type of learning and assessment activity in an

[3]According the National Center for Educational Statistics, even large majors such as computer science only have on the order of 40,000 graduates every year. Smaller majors might have single-digit thousands. A high-caliber but narrow school such as the Massachusetts's Institute of Technology (MIT) offers roughly 2,000 courses to cover primarily science and engineering education. If we assume similar numbers of courses across disciplines MIT does not offer, such as agriculture, education, medicine, or law, a complete set of courses to cover a broad university education would require about 10,000 courses. As of 2017, there are thousands of MOOCs.

economical fashion for the number of courses in MOOCs. It is also uncertain how systems like AutoTutor will integrate into the design of current MOOCs.

The written word can also be the focus of MOOC assessment through the grading of essays and other extended written work. Several MOOCs have used peer review to grade essays, where (as mentioned previously) students grade each others' work and give feedback (Balfour, 2013). However, peer review may be less useful for problems with a large expert-novice gap, where a substantial portion of the goals may be, for instance, to give feedback on how well students conform to good design practice or other professional conventions. In addition, many students prefer to have their assignments graded by an expert rather than a peer (who may even be less knowledgeable than they themselves are) (Luo et al., 2014). However, in the majority of cases, with a clear, well-designed, and relatively closed-ended rubric, and proper calibration, peer reviews can be more reliable than a single expert review for suitable assessments – and are definitely more scalable.

Another potential approach to grading student essays is to use auto-grading, based on NLP. NLP-based auto-grading has been used at scale in other domains, perhaps most notably in standardized examinations. A thorough review of key systems for automated essay scoring can be found in (Dikli, 2006). Reilly and colleagues (2014) report on the use of automated essay scoring in a pharmacy MOOC, finding good agreement between automated scoring and instructor scores. However, automated essay scoring remains controversial in the context of MOOCs. edX piloted this capacity and found that when appropriately used, it had very high quality results (Mitros et al., 2013). However, practical engineering constraints prevented this approach from out to large numbers of courses. In particular, for such algorithms to work, the instructor must first hand-grade around 100 submissions. These free-form text submissions are typically unavailable the first time a MOOC is run, creating a chicken-and-egg problem. In addition, while the system worked very well in the courses in which it was piloted it is difficult to predict whether it would work in all such contexts. If an auto-grading system failed in a MOOC, leaving thousands of essays to be hand-graded, such a failure could either be exceptionally expensive or highly problematic. Such problems are solvable, whether by integrating with peer grading, use in prototype courses to obtain initial models, using teaching assistants in the developing world, or other solutions, but no solutions has currently been developed to the level of being production-ready for thousands of courses.

Across approaches to assessment and instruction, one key lesson learned is that in MOOCs, usage of a feature is strongly tied to how easy it is for course authors to discover and use that feature. Course authors are willing to invest significant effort into making high-quality courses if they can figure out how to do so. For example, ITSs have a range of hinting functionality, including on-demand hints and so-called "bug messages" for incorrect answers. Such hints both enhance learning and have been shown to be a valuable component of more precise assessment, providing data beyond just student correctness that is predictive of long-term outcomes (Feng, Heffernan & Koedinger, 2009). The edX platform added simple authoring functionality for hinting in 2015, including hints in default template problems. As a result, as of this writing, around ⅔ of edX assessments have hint functionality, either as on-demand hints or "bug messages".

Assessment: Beyond Knowledge

Thus far in this chapter, we have discussed the assessment of students in MOOCs as if the only thing worth assessing is students' knowledge and skills. It is true that knowledge and skill have historically been the primary focus of assessment work, across contexts and domains, but they are hardly the only constructs that can be assessed, or the only constructs that should be assessed. MOOCs are well suited to assessing more complex skills, such as group work, creative problem solving, and leadership. MOOCs capture minute click-by-click student interactions, across a diverse range of subjects, with data for some learners across up to five years. These data have the potential to help us study and assess student progress across a sequence

of several complex and group projects, providing insights into the details of the social interactions and problem solving within those projects (Mitros et al., 2014). However, MOOCs have not yet reached their full potential in this area.

For instance, one of the key areas of research in the ITS community over the last decade has been the assessment of metacognition and self-regulated learning (SRL) skill. MOOC students are generally gifted high school students, college students, and adult learners, and even within that set, are disproportionately autodidacts. They can be hypothesized to have higher levels of SRL skills and metacognition than the more general population of learners whom ITSs traditionally target. The edX platform, while offering a linear default path through content, was designed to support SRL strategies by providing supplementary resources that students can choose to access, as well as multiple navigational elements for students to be able to monitor their learning, skip over material they know, or navigate back to material they did not adequately master[4]. However, evaluation of how well these design elements work is limited, and it is not yet known to what degree the edX design works as intended. In terms of assessing metacognition and SRL within MOOCs, some work has focused on the use of out-of-context questionnaires rather than recognizing SRL from behavior (Hood et al., 2015; Onah & Sinclair, 2016). However, it has been argued that this type of questionnaire does not capture key aspects of SRL (Winne & Baker, 2013). Other work has looked at whether students' navigation patterns in MOOCs follow the default linear path, but has not fully closed the loop from quantitative description to qualitative understanding (Guo et al., 2014). We do see that a significant number of students completing MOOCs skip over significant numbers of videos (Seaton et al., 2014), which suggests some different learning strategies are being applied.

In the context of ITSs, models have been developed that can recognize a range of SRL strategies, from unscaffolded self-explanation (Shih, Koedinger & Scheines, 2011) to help-seeking (Aleven et al., 2006). Models that can assess help-seeking skills have been used as the basis of automated support for SRL, leading to systems that produce enduring positive changes in students' help-seeking strategies (Roll et al., 2011). Even simple training in strategies for planning, monitoring, and knowledge elaboration can lead to better learning outcomes in laboratory studies (Azevedo & Cromley, 2004). However, simply recommending SRL strategies to MOOC learners does not appear to lead to benefits (Kizilcec et al., 2016). Overall, the best way of encouraging effective SRL strategies in MOOCs is an open question.

MOOCs offer several opportunities for measuring SRL at a behavioral level, including student use of discussion forums to use questions, and student activity in the face of incorrect answers within a knowledge assessment – after making a mistake, does the student give up? Try again immediately? Ask for help (perhaps on the discussion forum)? Re-watch the video? By modeling these behaviors, we may be able to assess SRL in MOOCs in the same rich fashion as has been achieved for ITSs.

Another area of assessment in MOOCs that goes beyond knowledge and skills, and which has received relatively more attention, is the assessment of student engagement. Inspired by student success systems in for-credit programs (Arnold & Pistilli, 2012), many researchers have attempted to identify the factors associated with a student not completing a MOOC and predict in advance whether or not (and when) a student will stop participating in a MOOC (Jiang et al., 2014; Kloft et al., 2014; Yang et al., 2013; Sharkey et al.,

[4]6.002x, the original edX course, was structured as learning sequences, composed of roughly a dozen elements each, replacing what would be lectures in traditional classrooms. These was linear navigation with back/forward buttons, but also a set of icons, one for each element of the sequence, with the icon indicating the type of element (e.g., problem vs. video), and tooltips describing what each element is about. Students, anecdotally, took multiple strategies. For example, some students would navigate to assessments and only watch videos if they had problems with those assessments. Videos had links to multiple additional means of presentation, such as textbook pages. Within the video, multiple speeds were available, for moving through the video more quickly or more slowly. In addition, a scrolling transcript allows students to read ahead and navigate to precise points in a video.

2014). Across studies, it appears that several forms of participation are associated with MOOC completion, including posting to discussion forums, reading discussion forums, completing assignments (somewhat tautologically, since completion of a MOOC is typically based on completing assignments with a sufficiently high grade), and watching videos (see review in Andres et al., in press).

A related area of research is in attempting to infer MOOC learners' emotions or sentiment. This is a well-established area in other types of online learning system (see the review in Baker & Ocumpaugh, 2014), with researchers developing, validating, and using models of affective states such as boredom, frustration, and engaged concentration/flow. Researchers have begun studying assessment of sentiment in MOOCs as well. For instance, Wen and colleagues (2014) use discussion forums data on Coursera to determine if students have positive or negative attitudes to a course's lectures, assignments, and peer assessments. Chaplot and colleagues (2015) show that student sentiment from discussion forums can be incorporated into models that infer whether a student will drop out of a course, leading to more accurate prediction of retention. However, there is not yet work to detect emotions in MOOCs beyond simply positive and negative sentiment; research into more complex emotion in MOOCs has thus far depended on self-report instruments (Dillon et al., 2016) rather than the automated detectors used with ITSs and other types of artificially intelligent learning software.

Finally, some researchers have begun to study how behavior within MOOCs can be predictive not just of completing the MOOC, but of students' career trajectories after the MOOC. Chen and colleagues (2016) find that many students learning in programming MOOCs take their knowledge beyond the MOOC, incorporating new programming knowledge into their publicly released software on gitHub. Wang and colleagues (2017) have determined that reading discussion forums is associated with submitting scientific papers in the field after course completion, but that posting is not associated with submitting papers, even for a MOOC where posting *is* associated with course completion. Assessing not just where a learner is today during a MOOC, but determining how it influences their future career, has the potential to help us better design MOOCs to positively impact students' long-term trajectories.

Assessments in GIFT to Drive MOOC Adaptation

As previously reviewed, assessment techniques applied across MOOCs varies based on the domain being instructed and the activities and exercises configured across lesson interactions. While many of the aforementioned assessments described are aimed at classifying performance states and comprehension levels, it is important to recognize the role these assessments can play in instructional management and remediation practices. With personalization and individualized course-flow serving as a recognized gap in the majority of current MOOC implementations, a current collaborative research effort involving Carnegie Mellon University, the University of Pennsylvania, and the US Army Research Laboratory (ARL) is investigating the utility of the Generalized Intelligent Framework for Tutoring (GIFT) for serving as a framework to structure MOOC content and lessons. In the enhanced MOOC this project is producing, configured assessments across the relevant MOOC-related activities will drive instructional management decisions at the run-time level based on GIFT's pedagogical configurations.

The effort is broken into two phases. The first phase of development focuses on making GIFT LTI compliant, for the purpose of interoperating with large-scale learning management system (LMS) sites like edX. This enables MOOC developers to reference GIFT-managed lessons within the structure and delivery of their course flow, along with the ability to receive data back following the completion of a GIFT lesson for performance tracking and accreditation purposes.

With the LTI component in place, the next phase involves configuring MOOC content into a set of lessons that adhere to the authoring standards and run-time schemas of GIFT. GIFT is unique because it provides

a domain-agnostic architecture that enables a course developer to build and sequence content within an instructional design theory that adheres to knowledge development and skill acquisition. GIFT's Engine for Management of Adaptive Pedagogy (EMAP) is a pedagogical model embedded within the 'Adaptive Courseflow' course object, with David Merrill's Component Display Theory (CDT) informing the design (Goldberg, Hoffman & Tarr; 2015; Merrill, 1994). When the EMAP is used, an author configures content for the delivery of "Rules" and "Examples" for each identified concept a lesson targets, followed by configuring two levels of assessment: 1) knowledge recall as it pertains to the declarative and procedural information and 2) skill and application assessment as captured within a set of practice events and/or scenarios.

The "Recall" and "Practice" components of the EMAP can support any number of MOOC related assessment activities, where the derived outcomes of the measures are used to drive what the learner experiences next. The decision involves letting the learner advance to the next configured interaction or remediation logic that is triggered, where the underperforming concepts or states are addressed through an intervention that targets the impasses or misconception identified. The utility of such an approach in a large-scale delivery scenario like an edX MOOC requires experimentation to determine impact and gauge overall effect.

Conclusion

In this chapter, we have briefly discussed the rich state of the art and the relatively more limited state of the practice of assessment in MOOCs. Although some MOOCs today offer calibrated peer assessments, automatically graded essays, step-by-step problem solving, and psychometrically based assessment, most MOOCs continue to base assessment on somewhat arbitrary sets of multiple-choice and fill-in-the-blank items. Furthermore, many of the technologies pioneered in ITSs and other types of artificially intelligent software, such as natural language dialogues, remain unavailable in the MOOC world, and significant barriers exist to bringing them to practice in such systems. Similarly, though there is some work to study metacognition, SRL, and sentiment in MOOCs, MOOC research in these areas has still not reached the level of sophistication seen other areas of educational research.

In some ways, this finding is not surprising. Although there has been an impressive quantity of research conducted on MOOCs, the history of MOOC research remains rather brief. Several of the most impressive demonstrations of the power of assessment in online learning have involved expensive, several-year research efforts. With time, MOOC assessment may reach the same peak in sophistication as ITSs. Having said that, the bigger challenge will be to roll out these benefits to the full diversity of existing MOOCs and use these forms of assessment to drive beneficial intervention. Even in the more mature field of ITSs, it has been challenging to develop interventions that take full advantage of the powerful forms of assessment now available. Solving this challenge in MOOCs will call on the collaboration of both assessment researchers and designers alike.

Acknowledgements

The research described herein has been sponsored by ARL. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred.

References

Aleven, V., Mclaren, B., Roll, I. & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*, *16*(2), 101–128.

Aleven, V., Sewall, J., Popescu, O., Xhakaj, F., Chand, D., Baker, R., Wang, Y., Siemens, G., Rosé, C., Gasevic, D. (2015) The Beginning of a Beautiful Friendship? Intelligent Tutoring Systems and MOOCs. *Proceedings of the 17th International Conference on Artificial Intelligence in Education*, 525–528.

Anderson, J. R., Corbett, A. T., Koedinger, K. R. & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The journal of the learning sciences*, *4*(2), 167–207.

Andres, J.M.L., Baker, R.S., Siemens, G., Gasevic, D., Spann, C.A. (in press) Replicating 21 Findings on Student Success in Online Learning. To appear in *Technology, Instruction, Cognition, and Learning.*

Arnold, K. E. & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267–270). ACM

Azevedo, R. & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia?. *Journal of educational psychology*, *96*(3), 523.

Bachelet, R., Zongo, D. & Bourelle, A. (2015). Does peer grading work? How to implement and improve it? Comparing instructor and peer assessment in MOOC GdP. In *European MOOCs Stakeholders Summit 2015*.

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review (tm). *Research & Practice in Assessment*, *8*.

Baker, R.S.J.d., Ocumpaugh, J. (2014) Interaction-Based Affect Detection in Educational Software. In R.A. Calvo, S.K. D'Mello, J. Gratch, A. Kappas (Eds.), *The Oxford Handbook of Affective Computing*. Oxford, UK: Oxford University Press.

Bloom, B. S. (1987). A Response to Slavin's Mastery Learning Reconsidered. *Review of Educational Research*, *57*(4), 507–8.

Champaign, J., Colvin, K. F., Liu, A., Fredericks, C., Seaton, D. & Pritchard, D. E. (2014). Correlating skill and improvement in 2 MOOCs with a student's time on tasks. In *Proceedings of the 1st ACM Conference on Learning@Scale* (pp. 11–20). ACM.

Chaplot, D. S., Rhim, E. & Kim, J. (2015). Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks. In *AIED Workshops*.

Chen, G., Davis, D., Lin, J., Hauff, C.,& Houben, G-J. (2016). Beyond the MOOC platform: Gaining Insights aboutLearners from the Social Web. In *Proceedings of the 8th ACM Conference on Web Science*, pp. 15--24, Hannover, Germany. WebSci '16, ACM.

Clarke-Midura, J., Mayrath, M. & Dede, C. (2012). Thinking outside the bubble: virtual performance assessments for measuring inquiry learning.

Cobos, R., Gil, S., Lareo, A., Vargas, F.A. (2016) Open-DLAs: an Open Dashboard for Learning Analytics. *Proceedings of the 3rd Annual Conference on Learning @ Scale*, 265–268.

Colvin, K. F., Champaign, J., Liu, A., Zhou, Q., Fredericks, C. & Pritchard, D. E. (2014). Learning in an introductory physics MOOC: All cohorts learn equally, including an on-campus class. *The International Review of Research in Open and Distributed Learning*, *15*(4).

Corbett, A. T. & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, *4*(4), 253–278.

Dernoncourt, F., Taylor, C., O'Reilly, U. M., Veeramachaneni, K., Wu, S., Do, C. & Halawa, S. (2013). MoocViz: A large scale, open access, collaborative, data analytics platform for MOOCs. In *NIPS Workshop on Data-Driven Education*.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1).

Dillon, J., Ambrose, G. A., Wanigasekara, N., Chetlur, M., Dey, P., Sengupta, B. & D'Mello, S. K. (2016). Student affect during learning with a MOOC. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 528–529). ACM.

doignon, J. P. & Falmagne, J. C. (2012). *Knowledge spaces*. Springer Science & Business Media.

Dudycha, A. L. & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, *58*(1), 116.

Feng, M., Heffernan, N. & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, *19*(3), 243–266.

Fredericks, C., Lopez, G., Shnayder, V., Rayyan, S. & Seaton, D. (2016, April). Instructor Dashboards In EdX. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 335–336). ACM.

Goldberg, B., Hoffman, M. & Tarr, R. (2015). Authoring Instructional Management Logic in GIFT Using the Engine for Management of Adaptive Pedagogy (EMAP). In R. Sottilare, A. Graesser, X. Hu & K. Brawner (Eds.), *Design Recommendations for Intelligent Tutoring Systems: Authoring Tools (Volume 3)*: US Army Research Laboratory.

Graesser, A. C., Chipman, P., Haynes, B. C. & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, *48*(4), 612–618.

Guo, P. J. & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 21–30). ACM.

Hood N., Littlejohn A. & Milligan C., (2015). Context Counts: how learners' contexts influence learning in a MOOC, Computers & Education.

Jiang, S., Williams, A., Schenke, K., Warschauer, M. & O'dowd, D. (2014, July). Predicting MOOC performance with week 1 behavior. In *Educational Data Mining 2014*.

Kizilcec, R. F., Pérez-Sanagustín, M. & Maldonado, J. J. (2016). Recommending self-regulated learning strategies does not improve performance in a MOOC. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 101–104). ACM.

Kloft, M., Stiehler, F., Zheng, Z. & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 60–65).

Leelawong, K. & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education*, *18*(3), 181–208.

Luo, H., Robinson, A. C. & Park, J. Y. (2014). Peer Grading in a MOOC: Reliability, Validity, and Perceived Effects. *Journal of Asynchronous Learning Networks*, *18*(2).

Merrill, M. D. (1994). *The descriptive component display theory*: Educational Technology Publications, Englewood Cliffs, NJ.

Mislevy, R. J. & Riconscente, M. M. (2006). Evidence-centered assessment design. *Handbook of test development*, 61–90.

Mitros, P. , et. al. (2013). Teaching Electronic Circuits Online: Lessons from MITx's 6.002x on edX. IEEE ISCAS.

Mitros, P., Agarwal, A., Paruchuri, V. (2014) Ubiquity symposium: MOOCs and technology to advance learning and learning research: assessment in digital at-scale learning environments. *Ubiquity*, 1–9.

Onah, Daniel F. O. and Sinclair, Jane (2016) Exploring learners' strategies of self-regulated learning abilities in a novel MOOC Platform : eLDa. In: 23rd Annual Conference of the Association for Learning Technology (ALT2016), University of Warwick, United Kingdom, 6–8 Sep 2016.

Pardos, Z., Bergner, Y., Seaton, D. & Pritchard, D. (2013, July). Adapting Bayesian knowledge tracing to a massive open online course in edx. In *Educational Data Mining 2013*.

Reilly, E. D., Stafford, R. E., Williams, K. M. & Corliss, S. B. (2014). Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *The International Review of Research in Open and Distributed Learning, 15(5).*

Roll, I., Aleven, V., McLaren, B. M. & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, *21*(2), 267–280.

Salden, R. J., Koedinger, K. R., Renkl, A., Aleven, V. & McLaren, B. M. (2010). Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review*, *22*(4), 379–392.

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P. & Pritchard, D. E. (2014). Who does what in a massive open online course?. *Communications of the ACM*, *57*(4), 58–65.

*S*harkey, M. & Sanders, R. (2014, October). A process for predicting MOOC attrition. *In Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 50–54).

Shih, B., Koedinger, K. R. & Scheines, R. (2011). A response time model for bottom-out hints as worked examples. *Handbook of educational data mining*, 201-212.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S.B., Ferguson, R., Duval, E., Verbert, K., Baker, R.S.J.d. (2011) Open Learning Analytics: an integrated & modularized platform: Proposal to design, implement and evaluate an open platform to integrate heterogeneous learning analytics techniques. Athabasca, Alberta, Canada: Society for Learning Analytics Research.

Viswanathan, R. (2012). Teaching and Learning through MOOC. *Frontiers of Language and Teaching*, *3*(1), 32-40.

Wang, Y.E., Baker, R., Paquette, L. (2017) Behavioral Predictors of MOOC Post-Course Development. *Proceedings of the Workshop on Integrated Learning Analytics of MOOC Post-Course Development*.

Wang, Y. & Heffernan, N. (2013). Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes. In *International Conference on Artificial Intelligence in Education* (pp. 181-188). Springer Berlin Heidelberg.

Wen, M., Yang, D. & Rose, C. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? *Proceedings of the International Conference on Educational Data Mining*.

Winne, P. H. & Baker, R. S. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *JEDM-Journal of Educational Data Mining*, *5*(1), 1-8.

Yang, D., Sinha, T., Adamson, D. & Rosé, C. P. (2013, December). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop* (Vol. 11, p. 14).

SECTION II

EVIDENCE-CENTERED DESIGN AND DATA MINING

Dr. Arthur Graesser, Ed.

CHAPTER 9 – Evidence Centered Design and Data-Driven Assessment

Arthur C. Graesser
University of Memphis

Core Ideas

The chapters in this section are substantially pushing the envelope beyond classical psychometric assessment of psychological constructs, such as those that analyze multiple-choice tests with item response theory and Rasch models. One important advance is that these approaches can analyze open-ended behavior, such as essays, answers to questions in natural language, conversation, collaborative interaction, problem solving, actions during scientific inquiry, and design of artifacts. This requires a fine-grained analysis of log files that record the stream of actions, events, processes, timing, and sometimes even physiological and emotional responses.

A second important advance in these approaches is that they incorporate contemporary breakthroughs in computational sciences, such as artificial intelligence, computational linguistics, information retrieval, machine learning, data mining, and multichannel multisensory signal processing. These methods are needed to interpret the complex log file data automatically and to map data patterns to performance indicators and psychological constructs that make sense.

A third important advance is that the models attempt to achieve a careful balance between theory and data-driven discovery. Patterns of data in the log files sometimes match expected theoretical constructs that can be specified *a priori*. For example, taking the initiative in a conversation is manifested in a person when the person asks questions and introduces new topics. Or systematic inquiry is manifested by a person who manipulates one variable at a time in a simulation environment. On the other hand, some patterns of data are discovered through data mining and machine learning analyses that eventually are integrated with an ever-evolving theoretical framework. For example, there is a difference between 1) a person who asks for hints in an intelligent tutoring system (ITS) to help the person understand a difficult concept and 2) a person who abuses a hint help facility by quickly and mechanically asking for hints to finish the problem without learning anything. Data mining methods have differentiated these two types of learners.

A fourth importance advance is that these methods attempt to integrate the computational sciences with rigorous quantitative psychometric models. The field of ITSs ignored psychometric methodologies for decades, whereas these chapters are attempting to break down the barriers. Both fields are destined to grow from this integration.

The obvious implication for the Generalized Intelligent Framework for Tutoring (GIFT) is that these approaches to assessment need to be incorporated in the GIFT architecture. These approaches allow automated assessments of a variety of psychological constructs that can be stored in the student model and used in recommender systems of intelligent tutoring systems. The profile of knowledge, skills, and abilities in the student model can be expanded by the world of psychometric assessment. The empirical results of ITSs can also be appreciated by a wider community of researchers and other stakeholders in the learning sciences who have high standards of assessment methodology.

Individual Chapters

The chapter by Mislevy and Yan describes the fundamentals of evidence-centered design (ECD) and how ECD would be applied to ITSs. ECD "embodies an evidentiary argument to reason from what students say or do in particular task situations, to claims about what they know or can do more broadly". For example, they have analyzed a Hydrive ITS that simulates many of the important cognitive and contextual features of troubleshooting the hydraulics systems of a F-15 aircraft; this simulation involves the operation of flight controls, landing gear, the canopy, the jet fuel starter, and aerial refueling. ECD has a multilayer process of creating an assessment of both specific and general proficiencies from the stream of behaviors, events, and contextual features in the log file. One of these layers is the conceptual assessment framework, which is the main blueprint for the assessment that interrelates the student model, the evidence model, and the task model. ECD assessment is currently the dominant model of assessment for complex open-ended tasks, such as interacting with computer simulation trainers, games, collaborative interaction, design of artifacts, and hopefully more ITS in the future.

The chapter by Zapata-Rivera, Brawner, Jackson, and Katz uses ECD in conversation based assessments with conversational agents (virtual characters). That is, students interact with one or more agents using natural language (i.e., speech or written responses) or menu-based predefined responses. The paths of conversational interactions in the conversation space are linked to particular psychological constructs that are being assessed. Conversation-based assessments have been applied to a variety of skills, including scientific inquiry skills, literacy, and mathematics. This chapter shows how the conversation-based assessment scenarios can be reused on new domain content, such as transferring from volcanoes to the weather. This chapter also describes how an evidence model could be incorporated in the GIFT architecture.

The chapter by LaMar, Baker, and Greiff assesses inquiry skills in problem solving and the practice of science. Traditional, simple, static assessment items cannot capture inquiry processes, so complex, interactive, dynamic tasks are needed. However, there is a challenge in identifying the inquiry skills manifested in the task on the basis of log data. The chapter describes three different methods for assessing inquiry skills. One uses theoretically defined features, such as the strategy of varying one thing at a time. The second uses data mining with machine learning to allow discovery of unexpected strategies. The third uses generative process models to compare student actions to agents that probabilistically implement specific inquiry strategies. The chapter describes the advantages, disadvantages, and appropriate assessment tasks for each method.

The chapter by Rus, Olney, Foltz, and Hu provides an overview of the opportunities, challenges, and state-of-the-art solutions in the area of automated assessment of learner generated natural language responses. Advances in computational linguistics have been incorporated in ITSs that have tutorial dialogue, automated scoring of essays, scoring of verbal answers to questions, and assessments of whether students' verbal contributions are semantically similar to expected good or bad answers. The chapter discusses computational approaches to interpreting natural language that vary from shallow detection of linguistic features to deep componential semantic analysis. The chapter summarizes recent breakthroughs in automating natural language and discourse, as well as toolkits in computational linguistics and large corpora that are available for machine learning projects.

The chapter by Greiff, Gasevic, and von Davier discusses the complexities of analyzing the log files of rich open-ended tasks to be assessed. The authors argue that an interdisciplinary effort is essential. The three authors include a cognitive psychologist, a psychometrician, and a computer scientist. The chapter identifies the most prominent challenges that are encountered when analyzing the large amounts of computer-generated process data and when trying to discover informative relationships among features and patterns of data that unfold over time. They consider how ITSs can facilitate research on assessment in addition to learning.

CHAPTER 10 – Evidence-Centered Assessment Design and Probability-Based Inference to Support the Generalized Intelligent Framework for Tutoring (GIFT)

Robert J. Mislevy and Duanli Yan
Educational Testing Service

Introduction

Simulation-based assessments open the door to a new paradigm of learning that is characterized by interaction and adaptation. One of the applications using simulation-based assessment is intelligent tutoring systems (ITSs). For example, physicians can practice heart surgery, and technicians can interactively learn skills and be trained to troubleshoot and repair aircraft in simulation-based ITSs. The widespread use of simulations for learning and assessment, and the need for ITSs in the Generalized Intelligent Framework for Tutoring (GIFT) is increasing.

In this chapter, we present an overview of such a framework, namely evidence-centered assessment design (ECD). ECD's language and representations support ITS designers in projects across different domains, task types, and purposes. In particular, ECD has proved useful as the design framework for simulation-based assessments (Clarke-Midura, Code, Dede, Mayrath & Zap, 2012; Mislevy, 2013), game-based assessments (Mislevy et al., 2014), and ITSs (Shute, 2011). The Hydrive ITS for learning to troubleshoot the hydraulics of the F-15 aircraft (Mislevy & Gitomer, 1996) is used here to illustrate ideas.

This chapter also describes probability-based inference in complex networks of interdependent variables for assessment (Almond, Mislevy, Steinberg, Williamson & Yan, 2015; Woolf, 2009) and offers recommendations for developing an ITS to support GIFT.

Background

People often face similar difficulties in a wide variety of domains. Novices don't know what information is relevant, how to integrate information, and what to do next. However, humans routinely display some amazing capabilities by working with patterns: patterns of perceiving, patterns of thinking, and patterns of acting. Thus, we can assemble these patterns flexibly in real time once we are sufficiently practiced. Expert performance is made possible through the continual interaction between the external patterns and the internal neural patterns, when the external patterns structure peoples' interactions in situations, such as language and professional practices, and the internal neural patterns recognize, make meaning of, and act through these external patterns (Ericsson, Charness, Feltovich & Hoffman, 2006; Greeno & van de Sande, 2007).

People develop their internal cognitive patterns through experience, that is, by participating in activities structured around these patterns, discerning regularities, seeing what happens as they act. Through reflective practice starting from simplified situations – best with feedback and support from others – people build their capabilities and overcome the pervasive limitations that plague novices (Ericsson, Charness, Feltovich & Hoffman, 2006).

A simulation can emulate some features of actual situations. It may change size, speed up, slow down, or simplify some aspects of real-world situations. An assessment designer must determine what aspects of real situations to include in a simulation environment and what aspects to modify or omit. These design choices must be made based on the intended purposes of the simulation.

Designing simulations for *learning* requires focusing on the features of situations for the targeted knowledge and skills, at a level that is just beyond the capabilities of the test-takers. Designing simulations for *assessment* requires a focus on information the assessment needs about test-takers' knowledge and skill.

The designers need to consider both simulations for learning and for assessment when designing an ITS. The kinds of problems and situations in which one learns to think and act in a domain are the same kinds of problems and situations that provide evidence about these capabilities. The realm of assessment designers and psychometricians is how that evidence to be evoked, captured, interpreted, summarized, and reported. An assessment design framework coordinates the contributions from the domain experts and simulation designers.

Evidence-Centered Design (ECD)

An educational simulation assessment embodies an evidentiary argument to reason from what test-takers say or do in particular task situations, to claims about what they know or can do more broadly. Messick (1994) summarizes its form neatly:

> *[We] would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 16).*

The ECD framework lays out the structures that contribute to instantiating an assessment argument in operational processes (Figure 1). Common language and representations across different forms of assessment help developers structure their work, both conceptually and operationally. These encourage reusability such as design patterns for generating tasks and adaptable scoring procedures.

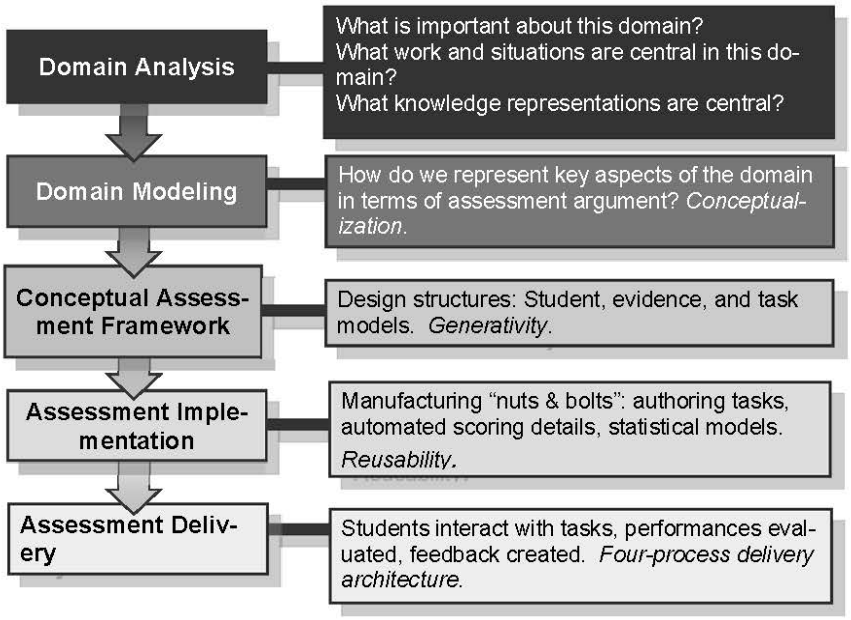


Figure 1. Layers of ECD, which include domain analysis, domain modeling, conceptual assessment framework (CAF), assessment implementation, and assessment delivery.

Domain Analysis

Domain analysis defines and documents the content or domains to be assessed. This is the relevant content knowledge, the ways people think and use it, and the kinds of situations they use it in. What do we know about progressions of thought or skills, patterns of common representations or errors? What knowledge, skills, goals, and tools are relevant? How do people interact with the physical environment, conceptual representations, and other people to accomplish goals? By what standards are these efforts judged in practice? Designers can consider the domain from a variety of perspectives including cognitive research, curricula, professional practice, expert input, standards, and current testing practices. This is information that developers draw upon when they design tasks, evaluation procedures, the nature of the test-takers' proficiencies they will report on and the interrelations among these.

Domain Modeling

Domain modeling is the process when designers organize information from domain analyses to articulate assessment arguments. ECD tools support this process with "claims and evidence" worksheets (Ewing, Packman, Hamen & Thurber, 2010), Toulmin diagrams for assessment arguments, and design patterns for constructing assessments around capabilities including design under constraints and model-based reasoning (Mislevy, Riconscente & Rutstein, 2009).

Optimal design requires careful consideration of which aspects are at issue for a given purpose. Messick's article "The Interplay of Evidence and Consequences in the Validation of Performance Assessments" (1994) remains the best source on how to think about what features should and should not be represented in simulations.

The Conceptual Assessment Framework (CAF)

The CAF is a blueprint for an assessment. It comprises the domain information, information about constraints and logistics, and models containing objects and specifications for operational aspects of work. These include 1) the creation of tasks, evaluation procedures, and statistical models, 2) delivery and operation of the assessment, and 3) analysis of data coming back from the field. While domain modeling emphasized the interconnections among key aspects of peoples' capabilities, situations, and behaviors, the CAF capitalizes on the separability of the objects that are used to instantiate an assessment.

Figure 2 is a schematic of the three central models in the CAF and objects they contain (Almond, Steinberg & Mislevy, 2002; Mislevy, Steinberg & Almond, 2003). The *student model* contains variables for expressing claims about targeted aspects of students' capabilities. Their number and character depends on the purpose of the assessment. For example, a single student-model variable can characterize students' overall proficiency in a domain of tasks for a certification decision, or a multidimensional student model can sort out patterns of proficiency from complex performances or provide more detailed feedback. In an ITS, the student-model variables are keyed to the nature and grain size of feedback and instruction that the ITS is intended to provide. This is illustrated in the Hydrive example.

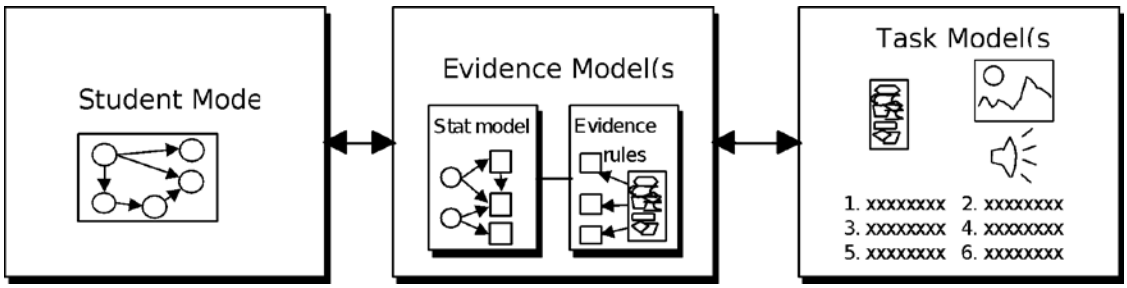


Figure 2. The central models of the CAF.

A *task model* formally describes the environment in which students say, do, or make something to produce evidence (Vendlinski, Baker & Niemi, 2008). In simulation tasks, more information is required than in fixed tasks because of the variety of actions a test-taker can take and the reactions of the simulation environment in response. For a simulation task, the initial status and transition rules of a finite state machine can be used for task model. (Gomaa, 2010). A key decision is how best, in what form, to capture students' performances, the *work product(s)*. For example, a work product can be a sequence of steps in an investigation, the locations of icons dragged into a diagram, or the final solution of a design problem. Task model variables indicate salient features of a situation, and are used in evaluating performances and extracting data concerning the situation in the assessment argument. In an interactive assessment, dynamic task model variables are determined as the performance unfolds.

An *evidence model* connects the student model and the task model. The *evidence rules* identify and evaluates the salient aspects of the work products, expressed as values of observable variables. Evidence rules are procedures such as rubrics for human scoring or algorithms for automated scoring procedures. Information in the observable variables is synthesized in the *stat model* or *measurement model* component. The simplest measurement models are classical test theory models, in which scores for salient features are added. Modular measurement models assemble more complicated models such as those of item response theory or Bayesian networks (Mislevy & Levy, 2007; Shute, 2011). It is more complicated but more useful when the structures of evidentiary relationships in complex tasks and multivariate student models are expressed in reusable measurement model fragments. It's important for task authors to create unique complex tasks but know ahead of time "how to score them" (Mislevy, Steinberg, Breyer, Johnson & Almond, 2002).

Assessment Implementation

The assessment implementation is about constructing and preparing the operational elements specified in the CAF. This includes authoring tasks, developing and finalizing rubrics or automated scoring rules, estimating the parameters in measurement models, and producing simulation states and transition rules. It is suggested to use common and compatible data structures to increase the reusability and interoperability, thus to bring down the costs of simulation-based assessment in the areas of task design, measurement models, authoring frameworks, and automated scoring (Chung, Baker, Delacruz, Bewley, Elmore & Seely, 2008; Frezzo, Behrens, Mislevy, West & DiCerbo, 2009).

Assessment Delivery: The Four-Process Architecture

The assessment delivery is where students interact with tasks, their performances are identified and evaluated, and feedback and reports are produced. In Almond, Steinberg, and Mislevy's (2002) four-process delivery architecture, the processes pass messages among themselves in a pattern determined by the test's purpose. The messages are either data objects specified in the CAF (e.g., parameters, stimulus materials) or are produced by the student or other processes in data structures that are specified in the CAF (e.g., work

products, values of observable variables). Again, using common language, data structures, and a partitioning of activities promote the reuse of objects and processes, and interoperability across projects and programs.

Figure 3 shows the four principal processes. The *activity selection process* selects a task or activity from the task library, or creates one according to what is known about the student or the situation. The *presentation process* is responsible for presenting the task to the student, managing the interaction, and capturing work products. Work products are passed to the *evidence identification process* for task-level scoring. This process evaluates work product using the methods specified in the evidence model. It sends values of observable variables to the *evidence accumulation process* for test-level scoring, which uses the stat or measurement models to summarize evidence about the student model variables and produce score reports. A simulation-based task can involve many interactions among the processes in the course of a performance.

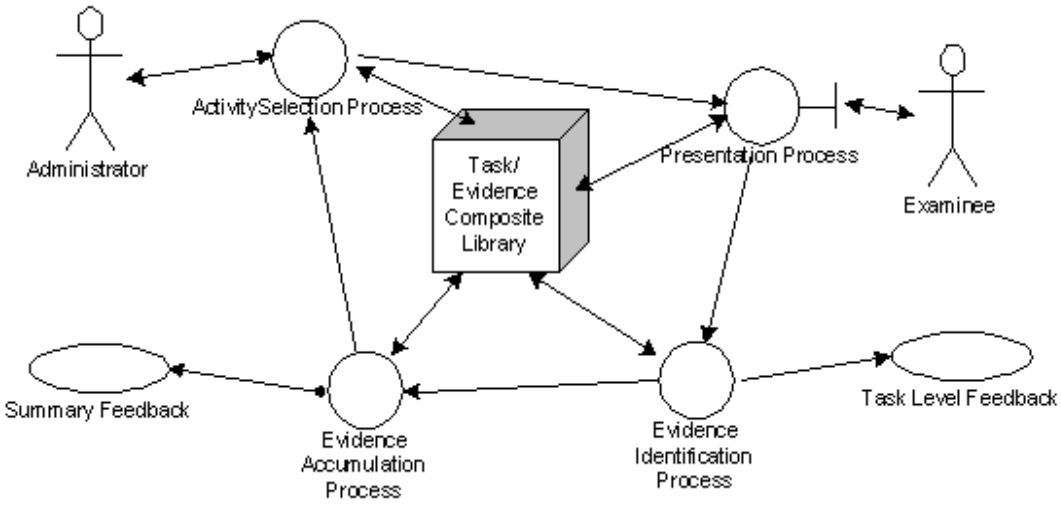


Figure 3. Processes in the assessment cycle.

Probability-Based Inference and Bayesian Networks for ITSs

Probability-Based Inference

Inference is reasoning from what we know and what we observe to explanations, conclusions, or predictions. We always reason in the presence of uncertainty. The information we work with is often incomplete, inconclusive, and amenable to more than one explanation (Schum, 1994). We attempt to establish the weight and coverage of evidence in what we observe because they inform the inferences and decisions we wish to make.

In probability-based inference, a "random variable" X is defined in terms of a collection of possible outcomes and a mapping from events to numbers that correspond to how likely they are to occur (probabilities). We denote $p(x)$ the mapping from a particular value x of X onto a probability. Basic probability structures can lead to consistent inference for very complex situations, such as game situations with unknown probabilities linked in complicated ways or with events whose probabilities depend on the outcomes of earlier observations (e.g., the probability of x given that another variable Z takes the value z, denoted $p(x|z)$). In

large-scale applications, these values can be verified empirically because the assessment can actually be used many times and the frequencies of various events tabulated.

Three kinds of reasoning play essential and interlocking roles in an ITS. Within an established framework of relationships among variables, *Deductive reasoning* flows from generals to particulars, that is, from causes to effects, from diseases to symptoms, from a student's knowledge and skills to observable behavior. *Inductive reasoning* flows in the opposite direction, from effects to possible causes, from symptoms to probable diseases, from students' solutions or patterns of solutions to likely configurations of knowledge and skill. Given outcomes, what state of affairs may have produced them? *Abductive reasoning* proceeds from observations to new hypotheses, new variables, or new relationships among variables.

In all ITSs, predications are made at each step of the tutoring process on some form of student modeling to guide tutor behavior. Inferences about a student's current skills, knowledge, and strategy usage can affect the type and pacing of problems, quality of feedback and instruction, and determination of when a student has completed some set of tutorial objectives. But we cannot directly observe what a student does and does not *know*; thus, we must infer, imperfectly, from what a student does and does not *do*.

Central to ITS development is the conception of the student model. A student model can fulfill at least three functions. First, given a set of instructional options, a student model can provide information to suggest which of the available choices on tasks or events is most appropriate for an individual at a given point in time (Ohlsson, 1987). Since an ITS can explicitly represent a domain of knowledge and task performance, it should be possible to design instruction at a level of cognitive complexity that facilitates successful performance and understanding (Kieras, 1988). Second, a student model enables prediction of the actions a student will take, given the characteristics of a particular problem state and what the system infers about the student's understanding (Ohlsson, 1987). With some understanding of students and problems, one can more accurately predict performance than if no model has been specified. When student actions conform to these predictions, it is an indication of the validity of the inferences about students made through the student model. Third, the student model enables the ITS to make claims about the competency of individuals with respect to their knowledge and various problem-solving abilities. These claims can be viewed as data summaries that can help the tutor make decisions about problem selection and exit criteria from a program of instruction, evaluate a person's status of knowledge and skills, and determine whether a person is likely to negotiate successfully a particular situation..

In practice, an ITS must work with specific actions that students take in specific situations. The student model mediates between the level of unique and unrepeatable observations, and the higher level of abstraction at which theory about the development of competence and the design of instruction takes place. The inferential task consists of two major parts: 1) establishing a framework for interpreting specific actions in terms suited to guide instruction and 2) characterizing the information these actions convey about variables in the student model. Using probability-based reasoning as a means for structuring the inferential task has many advantages. A distinctive feature of the approach is the differentiation between a model for a student's knowledge (i.e., values of variables in a student model space that encompasses key aspects of knowledge) and a model for an observer's state of knowledge about this student model (Mislevy, 1994).

Probability theory provides powerful mechanisms for explicating relationships, reasoning bidirectionally, criticizing and improving models, and handling evidentiary subtleties. The beliefs about the status and interrelationships of aspects of students' competences and actions can be approximated through variables modeling their interrelationships in a joint distribution.

Bayesian Inference Networks

In practical inference, we must reason inductively from interpreted actions to updated beliefs about the student's strategic knowledge. This is accomplished in probability-based reasoning by means of Bayes theorem. Let x be a variable whose probability distribution $p(x|z)$ depends on the variable z. Suppose also that prior to observing x, belief about the value of z can be expressed in terms of a probability distribution $p(z)$. For example, we may consider all possible values of z equally likely, or we may have an empirical distribution based on values observed in the past. Bayes Theorem says

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}, \tag{1}$$

where $p(x)$ is the expected value of x over all possible values of z – a normalizing constant required by the axiom that belief about z after having learned x must also be represented by a probability distribution that sums to one.

Efficient probability-based inference in complex networks of interdependent variables is an active topic in statistical research spurred by applications in such diverse areas as forecasting, pedigree analysis, troubleshooting, and medical diagnosis (e.g., Lauritzen & Spiegelhalter, 1988; Pearl, 1988). For an introduction to Bayes nets in cognitive diagnosis, see Mislevy (1995) and Martin and VanLehn (1993). Interest centers on obtaining the distributions of selected variables conditional on observed values of other variables, such as likely characteristics of offspring of selected animals given characteristics of their ancestors or probabilities of disease states given symptoms and test results.

A *recursive representation* of the joint distribution of a set of random variables $x_1,\ldots,x_n$ takes the form

$$p(x_1,\ldots,x_n) = p(x_n|x_{n-1},\ldots,x_1)p(x_{n-1}|x_{n-2},\ldots,x_1)\cdots p(x_2|x_1)p(x_1)$$

$$= \prod_{j=1}^{n} p(x_j|x_{j-1},\ldots,x_1), \tag{2}$$

where the term for $j = 1$ is defined as simply $p(x_1)$. A recursive representation can be written for any ordering of the variables, but one that exploits conditional independence relationships is useful because variables drop out of the conditioning lists. A graphical representation of Equation 2, or a directed acyclic graph (DAG), depicts each variable as a node; each variable has an arrow drawn to it from any variables on which it is directly dependent (its "parents"). Conditional independence corresponds to omitting arrows ("edges") from the DAG, thus simplifying the topology of the network.

The conditional independence relationships suggested by substantive theory play a central role in the topology of the network of interrelationships in a system of variables. If the topology is favorable, such calculations can be carried out efficiently through extended application of Bayes theorem even in very large systems. Table 1 is an example of conditional probabilities defining the relationship between two variables in a Bayes net. Each row is a conditional probability distribution, expressing the probability of certain classes of actions by a person with the level of strategic-knowledge proficiency in the row label at the left. There is a row for each possible value of the strategic knowledge student-model variable. Note the direction of reasoning represented here: If a person's strategic knowledge were such-and-such, then the probabilities of the different possible actions would be such-and-such.

Table 1. Numerical values of conditional probabilities of interpreted action sequences, given strategic knowledge.

	Conditional Probability of Interpreted Action Sequence			
Strategic Knowledge	**Serial Elimination**	**Redundant Action**	**Irrelevant Action**	**Remove and Replace**
Expert	0.75	0.10	0.05	0.10
Good	0.50	0.10	0.10	0.30
Okay	0.30	0.15	0.15	0.40
Weak	0.20	0.20	0.30	0.30

The following example shows how ECD ideas are applied to develop a Bayes net used in an ITS on a Hydrive tutoring/assessment system. The final form of a portion of the model, with probabilities set at initial values, is shown as Figure 4. It is comprised of student model variables representing states of belief about various aspects of trainees' proficiency and observable variables representing evaluations of action sequences they take in a simulated troubleshooting environment.

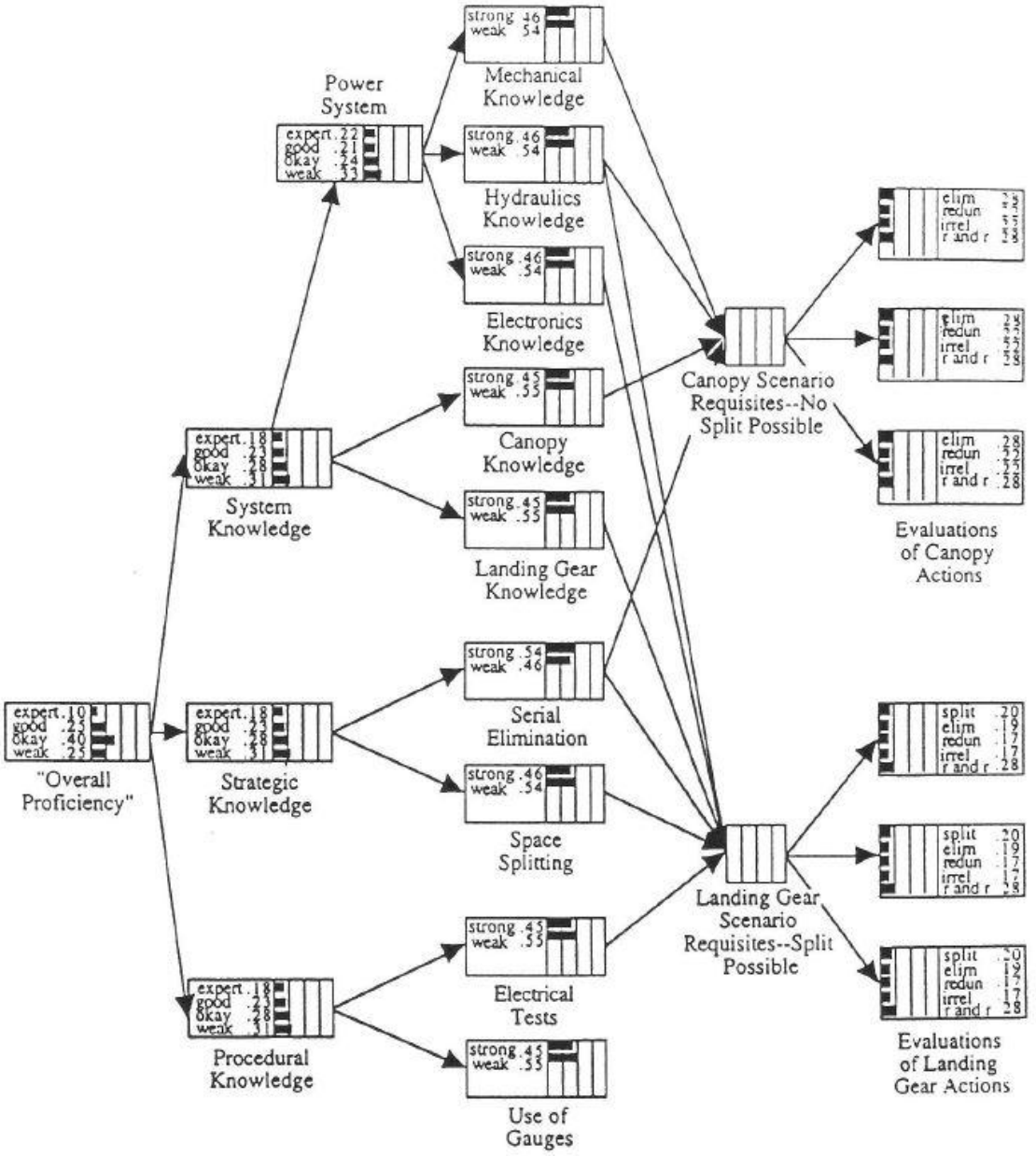


Figure 4. Conditional probabilities of interpreted action sequences in a canopy situation where space-splitting is not possible, given strategic knowledge.

An Example of an ITS: Hydrive

We now turn to discussing the implementation of probability-based reasoning in the Hydrive tutoring/assessment system for developing troubleshooting skills for the F-15 aircraft's hydraulics systems (Gitomer, Steinberg & Mislevy, 1995). This includes presenting an example of Bayesian inference networks for updating student models in ITSs. Specifically, we address issues encountered in defining variables, expressing their interrelationships, constructing conformable probability distributions, and carrying out inference, all illustrated in the context of Hydrive. We illustrate how probability-based inference can support generalized

claims about aspects of student proficiency through the combination of detailed epistemic analysis of particular actions within a system with probability-based inference. The psychology of learning in the domain and the instructional approach are seen to play crucial roles.

The hydraulics systems of the F-15 aircraft are involved in the operation of flight controls, landing gear, the canopy, the jet fuel starter, and aerial refueling. Hydrive is designed to simulate many of the important cognitive and contextual features of troubleshooting on the flightline. In the Hydrive ITS, a task starts with a video sequence in which a pilot, who is about to take off or has just landed, describes some aircraft malfunction to the hydraulics technician, for example, the rudders do not move during pre-flight checks. Hydrive's interface then offers the student several options, including performing troubleshooting procedures by accessing video images of aircraft components and acting on those components; reviewing online technical support materials, such as hierarchically organized schematic diagrams; and making their own instructional selections at any time during troubleshooting, in addition to or in place of instruction the system itself recommends. The state of the aircraft system, including the fault to be isolated and any changes brought about by user actions, is modeled by Hydrive's *system model*. The student's performance is monitored by evaluating how that student uses available information about the system to direct troubleshooting actions. Hydrive's *student model* is used to diagnose the quality of specific troubleshooting actions and characterize student understanding in terms of more general constructs such as knowledge of systems, strategies, and procedures that are associated with troubleshooting proficiency.

Defining Variables in Hydrive

Most real-world problems do not present themselves in terms of natural, ready-made "random variables" that are features of the world. Random variables are rather features of our representations of the patterns in terms of which we organize our thinking about the world (Shafer, 1988). To map any real-world situation into any formal reasoning framework, we must choose the level of detail at which variables will be defined, relationships will be modeled, and analyses will be carried out (Schum, 1994; Kadane & Schum, 1992).

"Strategic knowledge" is a clear abstraction that instructors use to summarize patterns of trainees' behavior, not just troubleshooting actions, but in their conversations, classroom activities, and interactions with instruction (see Pearl, 1988, p. 44, on the very human drive to invent such constructs to organize and explain our experience). For example, novice trainees use serial elimination informatively, but they tend to take space-splitting actions increasingly often as they gain competence and they take fewer redundant or irrelevant actions. We might therefore propose a variable called "strategic knowledge" for our student model, with possible values that represent increasing levels of expertise.

"Interpreted actions" constitute the interface between the individual-level variables of Hydrive's student model and the virtually unique sequences of actions that individual students take as they work through a problem. These are the lowest level of probability-based reasoning in Hydrive. The input to these variables corresponds to "observed data" for probabilistic reasoning, although they are actually fallible judgments from a rule-based parsing of students' actions. This is referred to as "virtual evidence" in the expert systems literature (e.g., Neapolitan, 1990, p. 230), and it highlights a source of uncertainty that we will have to take into account. The values of "interpreted action" variables are produced by Hydrive's *system model*, *action evaluator*, and *strategy interpreter*. A student's actions are evaluated in terms of the information they yield in light of the current state of the system model. The action evaluator calculates the effects on the problem area of a student's action sequence. The strategy interpreter makes rule-based inferences about the student's apparent strategy usage based on the nature and the span of problem area reduction obtained from the action evaluator.

The *system model* appears to the student as an explorable, testable aircraft system in which a failure has occurred. It is built around sets of components connected by inputs and outputs. Connections are expressed as pairs of components, the first producing an output that the second receives as an input, qualified by the type of power characterizing the connection. The system model processes the actions from the student and propagates sets of inputs and outputs throughout the system. A student activates the system model by providing input to the appropriate components, and then examines the results for any other component of the system. A student can move the landing gear handle down and then observe the operation of the landing gear. If the landing gear does not descend, the student may decide to observe the operation of other components to begin to isolate the failure.

The *action evaluator* considers every troubleshooting action from the student's point of view, in terms of the information it conveys about the problem area. When a student acts to supply power and inputs to the aircraft system, the effects of this input spread throughout the system model create explicit states in a subset of components: the active path, comprising the points from which input is required to initiate system function to its functionally terminal outputs and all the connections in between. The action evaluator updates its problem area as if the student correctly judged whether observations reveal normal or abnormal component states. If a student observes the output of a certain component that the system model "knows" is normal, then it is possible for the student to infer that all edges on the active path, up to and including the output edge, are functioning correctly and remove them from the problem area. If the student makes the correct interpretation and draws the appropriate inferences, then the problem areas that the student model and the student hold will in fact correspond and troubleshooting continues smoothly. But if the student decides that the observed component output was unexpected or abnormal, then a student may decide that all the edges in the active path would remain in the problem area and any others would be eliminated. If the problem area maintained by the student model begins to diverge significantly from the one present in the student's mind, then the student actions interpreted as irrelevant and redundant become more likely.

The *strategy interpreter* evaluates changes to the problem area, or the entire series of edges belonging to the system/subsystem where the problem occurs. As a student acts on the system model, the problem area is reduced because the results of action sequences, if correctly interpreted, eliminate elements as potential causes of the failure. If the student inspects any particular component, the system model will reveal a state that may or may not be expected from the student's perspective. Hydrive employed a relatively small number of strategy interpretation rules (~25) to characterize each troubleshooting action in terms of both the student and the best strategy.

Deductive and Inductive Reasoning

Consider a scenario near the end of a problem solution, where space-splitting is no longer an option. What are our expectations that a student at each level of strategic knowledge might perform, for action sequences interpreted as "serial elimination," 'redundant action," "irrelevant action," and "remove and replace"? Serial elimination is the best strategy available; remove and replace is useful but not efficient and both redundant and irrelevant actions are undesirable. Figure 5 shows the flow of deductive reasoning. Each subpanel depicts conditional probabilities of the various action categories, given level of strategic knowledge. These are values of $p(x|z)$ in Equation (1). Table 1 gives numerical values for this illustration. As the level of strategic knowledge increases, we see increasing likelihood for serial elimination and decreasing likelihood of redundant and irrelevant actions. It should be noted that experts sometimes make redundant moves, and novices make what are interpreted as expert moves (but not always for the right reasons).

Again, we must reason inductively in practice. Suppose that for a new student, before observing any performance, we start from initial beliefs of equal probability across the four possible values of strategic knowledge. That is, the prior probabilities $p(z)$ in Equation (1) are all 0.25. We then observe one action in

the scenario. Figure 6 illustrates the results of applying Equation (1) to calculate posterior probabilities $p(z|x)$, that is, updated beliefs about the test-taker's strategic knowledge that would be obtained if each of the possible evaluations were obtained. For example, if we observe an action interpreted as serial elimination and apply Bayes theorem, we obtain the results in the first panel of Figure 6. We maintain the direction of the arrow because this was the direction in which we specified conditional probabilities. Similar calculations would lead to the results in the other panels if we had observed any of the other possible interpretations.

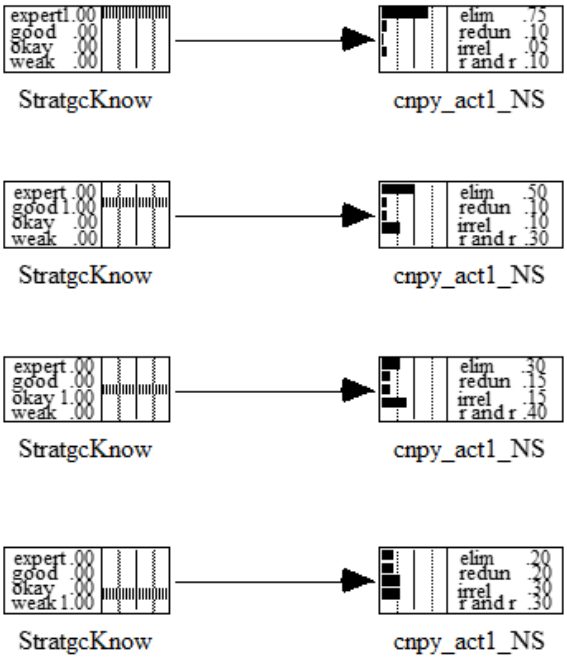


Figure 5. Conditional probabilities of interpreted action sequences in a canopy situation where space-splitting is not possible, given strategic knowledge.

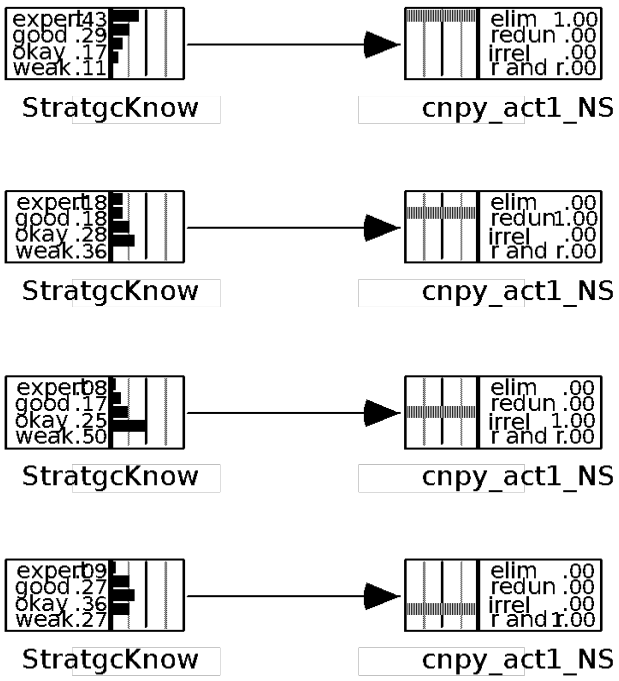


Figure 6. Updated probabilities for strategic knowledge, after observing an irrelevant action in a canopy situation where space-splitting is not possible.

Figure 4, introduced earlier, is a simplified Hydrive Bayesian inference network expressing the dependence relationships in simplified version of the inference network for the Hydrive student model. The direction of the arrows represents the deductive flow of reasoning for constructing probability distributions that incorporate the depicted dependence structure. A joint probability distribution for all these variables can be constructed by first assigning a probability distribution to each variable that has no parents (in this example, there is only one: "overall proficiency"); then for each successive variable, assigning a conditional probability distribution to its possible values for each possible combination of the values of its parents. Four groups of variables can be distinguished:

- The rightmost nodes are the "interpreted actions", based on the results of rule-driven epistemic analyses of students' actions in a given situation. There are two prototypical sets, each corresponding to an equivalence class of potential observables in a given scenario, and there are three members of the class are represented in both cases.

- The immediate parents of the interpreted action variables are the knowledge and strategy requirements that define the class in each case.

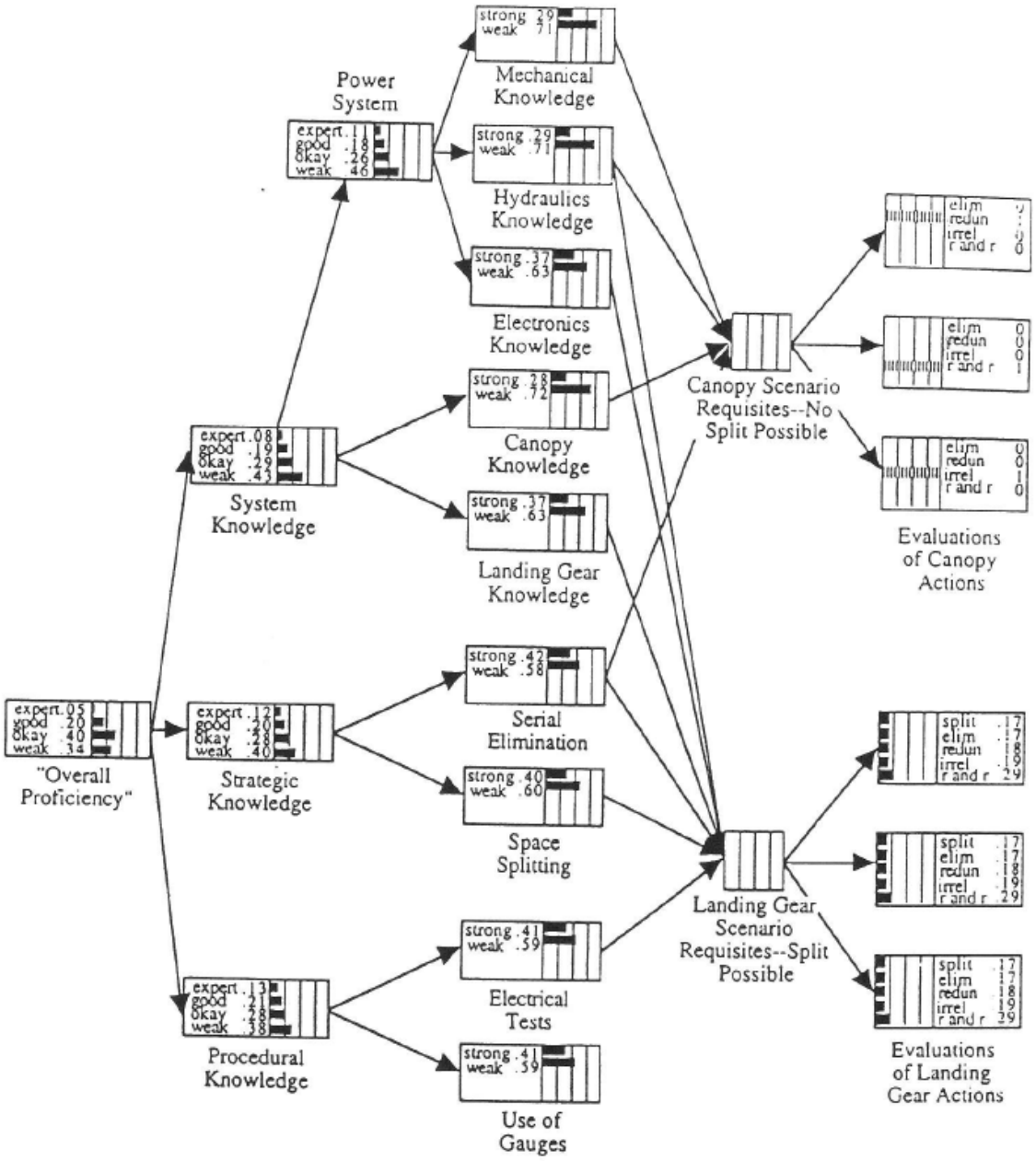
- The long column of variables in the middle concerns aspects of subsystem and strategic knowledge. These correspond to instructional options.

- The nodes to the left are summary characterizations of more generally construed proficiencies.

The equivalence classes of actions in this figure concern canopy situations in which space-splitting is not possible and landing gear situations in which space-splitting is possible. Figure 7 depicts belief after observing one redundant and one irrelevant action (both ineffectual troubleshooting moves) and one remove-and-replace (serviceable but inefficient) in three separate situations from the canopy/no-split class. Serial elimination would have been the best strategy in this case and is most likely to occur when the student has strong knowledge of this strategy and all relevant subsystems. Remove-and-replace is more likely when a student possesses some subsystem knowledge but lacks familiarity with serial elimination. Weak subsystem knowledge increases chances of irrelevant and redundant actions.

Subsystem and *strategy variables* are used to summarize tendencies in interpreted behaviors at a level addressed by instruction and disambiguate patterns of actions in light of the fact that inexpert actions can have several causes. These student-model variables are particularly salient in Hydrive, in that these variables correspond to companion modules of instruction and practice on particular aspects of systems, strategies, and tactics. When observations over a course of actions lead to a belief that a student may be weak in one or more of these areas, Hydrive recommends the training module(s) to the student. A contrasting ITS that also employed Bayesian networks but was constructed at a finer grain-size was the Online/Offline Assessment of Expertise (OLEA) tutor for kinematics (Martin & VanLehn, 1993). Its student-model variables were at the level of rules in a production system for mastering the course material. Instructional decisions and feedback were cast at this level, from this perspective.

Figures 7, 8, and 9 show the state of belief that would result after observing different sets of actions in situations involving the landing gear in which space-splitting is possible. To begin, suppose we first observe three inexpert actions concerning the canopy subsystem. The resulting updated Bayes net is shown in Figure 7. Belief for the student-model variables is shifted toward lower values for serial elimination and for all subsystem variables directly involved in the situation (mechanical, hydraulic, and canopy knowledge). Any or all could be a problem, since all are required for high likelihoods for expert actions. Variables for subsystems not directly involved in these situations are also lower, because to varying extents, students familiar with one subsystem tend to be familiar with others, and, to a lesser extent, students familiar with subsystems tend to be familiar with troubleshooting strategies. These relationships are expressed by means of the more *generalized system* and *strategy knowledge* variables at the left of the figure. These variables serve to exploit the indirect information about aspects of knowledge not directly tapped and summarize broadly construed aspects of proficiency for evaluation and problem selection.



Note: Bars represent probabilities, summing to one for all the possible values of a variable. A shaded bar extending the full width of a node represents certainty, due to having observed the value of that variable; i.e., a student's actual responses to tasks.

Figure 7. Status of a student after observation of three inexpert actions in canopy situations.

Suppose we then observed three additional *inexpert* action sequences dealing with the landing gear subsystem. Entering these findings and updating beliefs about student-model variables produces the beliefs in Figure 8. The status on all subsystem and strategy variables is further downgraded and is reflected in the more generalized summary variables.

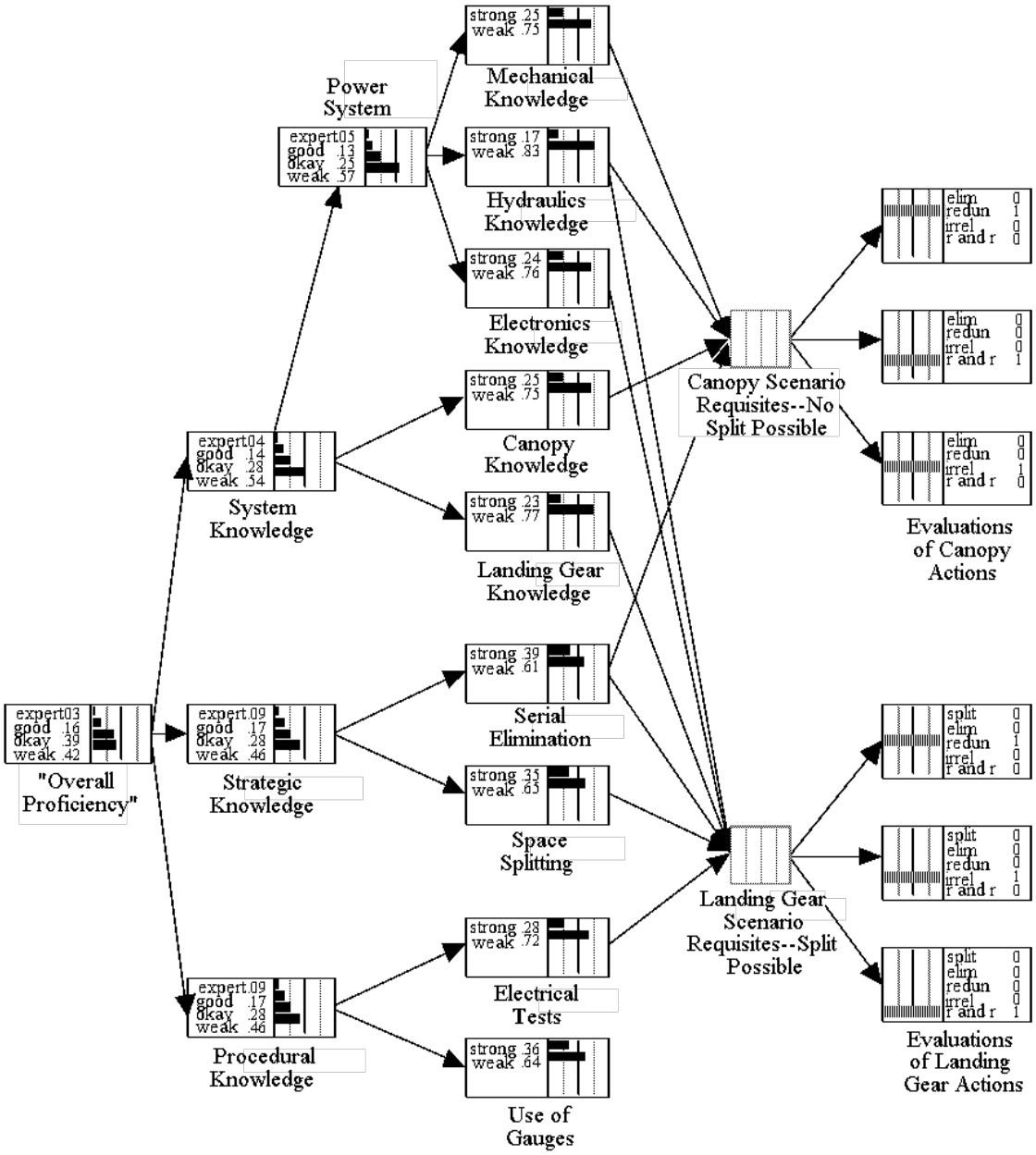
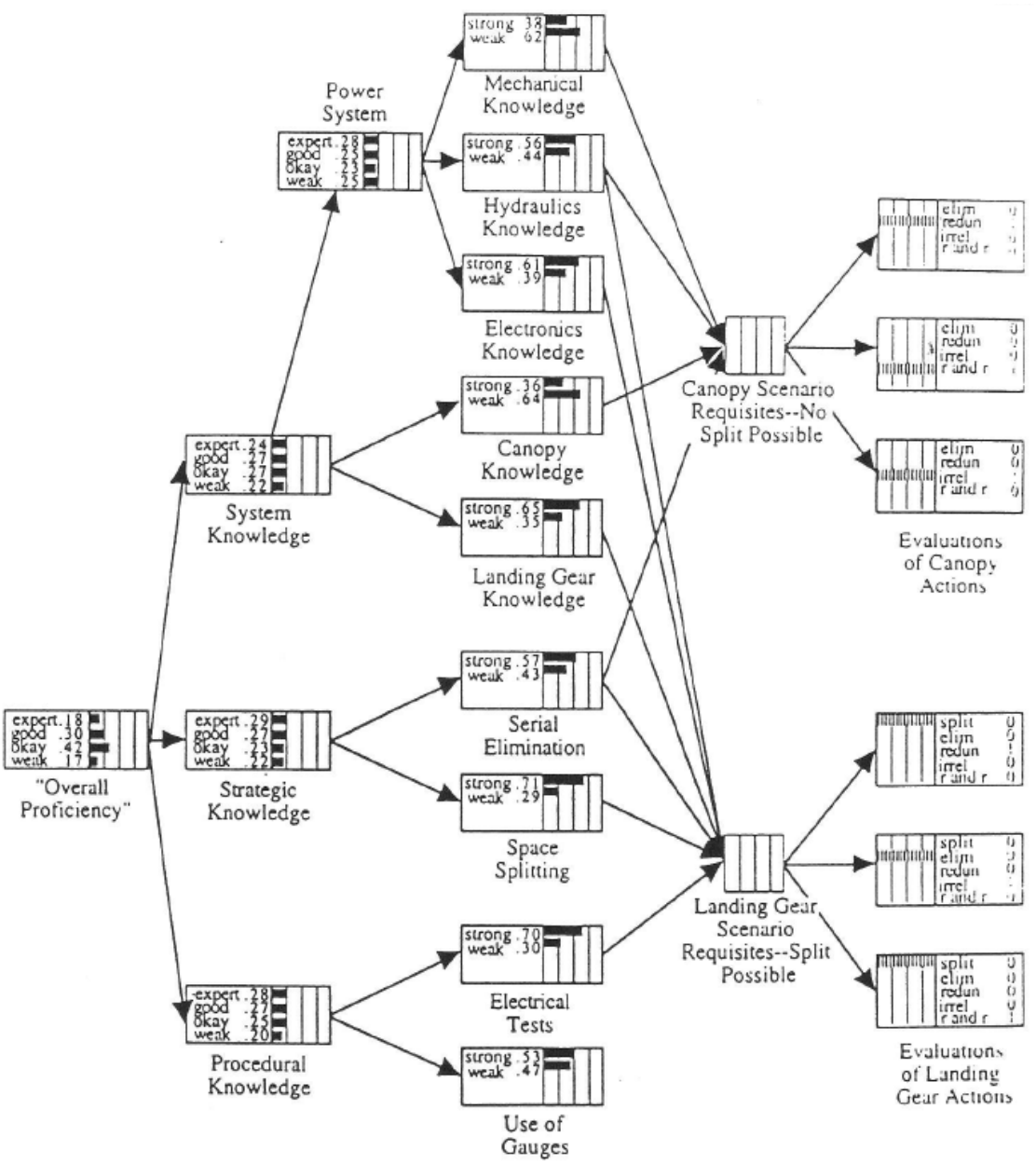


Figure 8. Status of a student after observation of three in-expert actions in both canopy situations and landing-gear situations.

On the other hand, suppose that after the initial three inexpert canopy actions, we then observed three more expert actions in landing-gear situations: two space-splits and one serial elimination. The updated beliefs in this case, depicted in Figure 9, show that belief about strategic skill would increase, as would beliefs

about subsystems involved in the landing-gear situations. Problems in mechanical and canopy subsystem knowledge are the most plausible explanations of the three inexpert canopy situation actions.



Note: Bars represent probabilities, summing to one for all the possible values of a variable. A shaded bar extending the full width of a node represents certainty, due to having observed the value of that variable; i.e., a student's actual responses to tasks.

Figure 9. Status of a student after observation of three in-expert actions in canopy situations and three more expert actions in landing-gear situations.

Modeling Learning Effects

The fundamental objective of an ITS is to help students *change* over time and improve their proficiencies. The preceding discussion and examples concerned updating belief about a static student model. That is, even though observations are obtained sequentially over time, they are presumed to simply provide additional information about values of student-model variables that remain constant over time. Most of our work have concentrated on modeling proficiencies within self-contained problem exercises. There are two other reasons for modifying belief about student-model variables: changes due to explicit instruction and changes due to implicit learning. In either case, the requirement under a probabilistic approach is to do so in a manner that maintains coherency. The approach described below accomplishes this end without requiring a full-blown dynamic model to be constructed and maintained.

Updating Based on Direct Instruction

Although Hydrive's system model functions as a discovery world for system and procedural understanding from the student's point of view, its student modeling components' evaluations are based on an implicit strategic goal structure observed in expert troubleshooting. This structure is made explicit in Hydrive's instruction. The student is given great latitude in pursuing the problem solution. Prompts or reminders (i.e., diagnostics) are given only when a student action constitutes an important violation of the rules associated with the strategic goal structure. Hydrive recommends direct instruction only when accumulating information across scenarios shifts belief (e.g., knowledge of a subsystem or strategy) sufficiently downward to merit more specifically focused feedback, review, and exercises. The student is free to follow this recommendation, choose different instruction, or continue troubleshooting without any instruction.

Such directed instruction can be expected to change students' understanding. For presumably static student-model variables, updating beliefs involved entering findings for the interpreted actions and propagating their implications upward through the network. For changes in student-model variables, updating beliefs involves direct manipulation of them. This implies that they are propagated both upward to related aspects of knowledge and downward to revise expectations for future actions. The degree of change is based on the student's performance on the exercises that accompany the instruction. This can be modeled in a small standalone Bayesian inference network that embodies a Markov process for change, while incorporating our uncertainty about the exact value of the student-model variable of interest. Both the probability distribution before instruction and the outcome of the instructional exercises are entered into the network. The modeled beliefs after instruction are output (Figure 10). Given the level before instruction and performance in exercises (e.g., Table 2), the conditional probabilities of student's level of competence after instruction may be refined over time, starting with expert opinion and limited experience then honed with the results of accumulating experience.

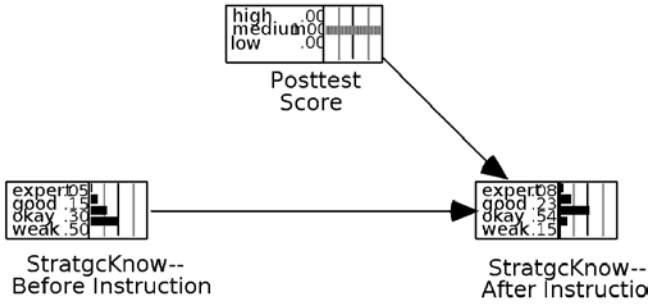


Figure 10. A Markov framework for direct updating of belief about strategic knowledge given results on a posttest score in an instructional module.

Table 2. Conditional probabilities of strategic knowledge after instruction, given probabilities before instruction and posttest performance.

Status before Instruction	Posttest Performance	Conditional Probability of Status after Instruction			
		Expert	Good	Okay	Weak
Expert	High	1.00	.00	.00	.00
	Medium	.95	.05	.00	.00
	Low	.90	.10	.00	.00
Good	High	.70	.30	.00	.00
	Medium	.20	.75	.05	.00
	Low	.05	.85	.10	.00
Okay	High	.20	.70	.10	.00
	Medium	.00	.30	.70	.00
	Low	.00	.15	.80	.05
Weak	High	.05	.55	.40	.00
	Medium	.00	.05	.65	.30
	Low	.00	.00	.20	.80

Updating Based on Learning While Problem Solving

Students can be expected to improve their troubleshooting skills as a result of practicing them and thinking through the problems, likely in increments throughout any given problem. Kimball (1982) employed an expedient in a calculus tutor: revisions to belief associated with implicit learning are effectuated only between problem boundaries. Kimball's tutor, like John Anderson's LISP tutor (Anderson & Reiser, 1985), revises belief in a manner consistent with probability axioms through an explicit learning model, a la Estes (1950). That is, a particular functional form for change is presumed, and degree of learning must also be assumed or estimated. We employ a more conservative and less model-bound approach, which accommodates student's learning by "forgetting" model or gradually discounting information from past as opposed to a "learning" model.

The basic idea is to enter each problem with student-model variable distributions that generally agree with the final values from the previous problem as to direction and central tendency, but are more diffuse and thus easier to change in light of new actions driven by possibly different (presumably improved) values. Levy (2014) describes dynamic Bayes nets to accomplish this blend of uncertainty due to unknown values of student-model variables and their change over time. Two simpler heuristic strategies are down-weighting the influence of actions as they recede in time and between problem sessions, mixing posterior distributions with noninformative distributions and propagating the revised versions through the network.

Recommendations

In GIFT, it is important to design a generalized ITS assessment. ECD provides a framework for evidentiary based assessment for ITS. Probability-based inference using Bayesian networks provide powerful machinery for coherent reasoning about complex and subtle interrelationships. This is achieved to the extent that one can capture, within its framework, the key aspects of a real-world situation (what is important, how important things are related, and what one sees or knows tells about what one doesn't see or doesn't know). If this can be accomplished, advantages both conceptual and practical accrue. A Bayes net built around the

generating principles of the domain makes interrelationships explicit and public, so one can not only monitor what one believes, but communicate why one believes it. A model can be refined over time in light of new information, as when originally subjective conditional probability specifications are updated in light of accumulating data. When able to calculate predictive distributions of any subset of variables given values of any others, one can investigate both deductive and inductive implications of a modeled structure, using hypothetical data to check for fidelity to what one believes or real data for fidelity to what one observes (see Crawford, 2014, Levy, 2006, Spiegelhalter, Dawid, Lauritzen & Cowell, 1993, and Williamson, Almond & Mislevy, 2000, on model-checking tools for complex Bayesian networks).

From this perspective, it may not be necessary or even desirable to attempt to exhaustively build all possible conjectures into one all-encompassing network. Shafer (1976, 1988) points out that in many inferential problems, frames of discernment often evolve over time as we accumulate evidence. We add possibilities, refine others, and abandon still others. Frameworks of probability-based reasoning aid our understanding of how available information informs current thinking, without claiming finality or "truth" at any stage; rule-based, inductive, and intuitive reasoning aid our construction and improvement of those frameworks.

Two areas of attention are germane in an ITS: 1) understanding about principles of domain and how people learn those principles in order to structure the student model efficaciously and 2) explicating what we need to see and how to interpret it in light of students' possible understandings in order to structure observable variables and their relationship to student-model variables. In Hydrive, we employed rule-based interpretations to identify critical features from a stream of relatively unstructured observations; Shafer (1987) sees the need for an associative memory mechanism for this purpose, strengthening the analogy to human perception.

This chapter has touched on three subjects: designing simulations, simulation-based learning, and simulation-based assessment. The three topics overlap substantially. For simulation-based learning, attention is on the cognitive patterns and activity patterns that people develop to be able to perform effectively in the relevant situations. Optimizing the evidentiary value of simulation-based performances requires going back to the design of the simulator and the situations with a psychometrician's perspective. Fewer options and more constrained situations may be less effective for learning but more effective for focusing test-takers' actions on key aspects of cognitive structures or activity structures. Requiring an explicit work product may slow working through a simulation, but it can make some valuable evidence about unobservable thinking manifest. Simply having lots of data, for example, gigabytes of time-stamped mouse clicks and key strokes, may not provide much evidence at all if it is not about salient actions in relevant situations.

This chapter has focused on the evidentiary-argument considerations that go into these design decisions using ECD. It does not have the space to discuss the importance of psychometric methods for modeling the data. It highlights assessment-related issues that the design team can understand and should be responsive to. It will require more esoteric methods from the psychometrician's toolbox to see exactly how different choices affect the evidentiary value of data, decision accuracy, or instructional effectiveness. A long-standing lesson applies: "To design a rich simulation environment to collect data without consideration of how the data will be evaluated, and hoping psychometricians will somehow 'figure out how to score it,' is a bad way to build assessments" (Mislevy, 2013). Data mining techniques are limited by the strength of the relationship between the information in the data and its connection to the targeted inferences (Mislevy, Behrens, DiCerbo & Levy, 2012). Close collaboration and interaction from the very start of the design process is preferable. Collaboration is needed among 1) users, who understand the purposes of the assessment intended; 2) domain experts, who know about the nature of the knowledge and skills, the situations in which they are used, and what test-takers do that provides evidence; 3) psychometricians, who know about the range of situations in which they can model data and examine its evidentiary value; and 4) software designers, who build the infrastructure to bring the assessment to life.

In particular, in using simulation as the basis of assessments like Hydrive, all of the complex considerations that go into designing simulation for learning are now joined by the equally challenging considerations that go into designing assessments. It is difficult enough to become an expert in either area, let alone both. The way forward builds on what has been learned from the rapid developments and proven successes of the various kinds of simulation in ITSs and coached practice systems. Grounding well-targeted exemplars such as Hydrive simulations in an assessment design framework shows how to integrate the deep principles of domains, learning, and assessment in simulation-based assessments. The concepts and representations of ECD help by explicitly building the principles of assessment design into the work, enhancing at once its efficiency, validity, and transferability.

A key idea is akin to Vygotsky's (1978) "zone of proximal development" (ZPD): Given where a learner is currently, what aspects of situations will best solidify skills, add the next layer of complexity, or develop new variations of a familiar theme, with the support of a teacher or software or appropriate structuring of the challenge? The features of situations that are critical for interaction must be in the simulation. That is, the means for the student to act on the system (its affordances) must be included, and the system must react to student's actions in ways that reflect the underlying principles of the system. Features to omit are those that add irrelevant complexity or require too much knowledge or skill that is not central. A high-fidelity system that models the way experts view problems may not be accessible to beginning students. Complexity can thus be introduced in stages to maximize effective learning. To help students learn optimally, a simulation system for learning can provide opportunities to slow down or stop action to reflect on what is important in a situation, what to do next, or why something happened.

References

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Williamson, D.M. & Yan, D. (2015). *Bayesian networks in educational assessment.* New York: Springer-Verlag.

Almond, R.G., Steinberg, L.S., Mislevy, R.J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5). Retrieved May 31, 2011 from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml.

Anderson, J.R. & Reiser, B.J. (1985). The LISP tutor. *Byte,* 10, 159–175.

Chung, G.K., Baker, E.L., Delacruz, G.C., Bewley, W.L., Elmore, J., Seely, B. (2008). A computational approach to authoring problem-solving assessments. In E.L. Baker, J. Dickieson, W. Wulfeck & H.F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 289–307). Mahwah, NJ: Erlbaum.

Clarke-Midura, J., Code, J., Dede, C., Mayrath, M. & Zap, N. (2012). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In D. Robinson, J. Clarke-Midura & M. Mayrath (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research.* Charlotte, NC: Information Age.

Crawford, A. (2014). *Posterior predictive model checking in Bayesian networks.* Doctoral dissertation, Arizona State University.

Ericsson, A.K., Charness, N., Feltovich, P., Hoffman, R.R.: (Eds.) (2006). *Cambridge handbook on expertise and expert performance.* Cambridge, UK: Cambridge University Press.

Estes, W.K. (1950). Toward a statistical theory of learning. *Psychological Review, 57,* 94–107.

Ewing, M., Packman, S., Hamen, C., Thurber, A.C. (2010). Representing targets of measurement within evidence-centered design. *Applied Measurement in Education, 23,* 325–341.

Frezzo, D.C., Behrens, J.T., Mislevy, R.J., West, P., DiCerbo, K.E. (2009). Psychometric and evidentiary approaches to simulation assessment in Packet Tracer software. *ICNS '09: Proceedings of the Fifth International Conference on Networking and Services* (pp. 555–560). Washington, DC: IEEE Computer Society.

Gitomer, D.H., Steinberg, L.S. & Mislevy, R.J. (1995). Diagnostic assessment of trouble-shooting skill in an intelligent tutoring system. In P. Nichols, S. Chipman & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 73–101). Hillsdale, NJ: Erlbaum.

Gomaa, H. (2010). *Software modeling and design.* Cambridge: Cambridge University Press.

Greeno J.G. & van de Sande, C. (2007). Perspectival understanding of conceptions and conceptual growth in interaction. *Educational Psychologist, 42,* 9–23.

Kadane, J.B. & Schum, D.A. (1992). Opinions in dispute: the Sacco-Vanzetti case. In J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith (Eds.), *Bayesian Statistics 4* (pp. 267–287). Oxford, U.K.: Oxford University Press.

Kieras, D.E. (1988). What mental model should be taught: Choosing instructional content for complex engineered systems. In M.J. Psotka, L.D. Massey & S.A. Mutter (Eds.), *Intelligent tutoring systems: Lessons learned* (pp. 85–111). Hillsdale, NJ: Lawrence Erlbaum.

Kimball, R. (1982). A self-improving tutor for symbolic integration. In D. Sleeman & J.S. Brown (Eds.), *Intelligent tutoring systems.* London: Academic Press.

Lauritzen, S.L. & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B, 50*, 157–224.

Levy, R. (2006). *Posterior predictive model checking for multidimensionality in item response theory and Bayesian networks.* Doctoral dissertation, University of Maryland.

Levy, R. (2014). *Dynamic Bayesian network modeling of game based diagnostic assessments* (CSE Report No. 837). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, Student Testing (CRESST). Retrieved from http://www.cse.ucla.edu/products/reports/R837.pdf.

Luecht, R.M. (2009). Adaptive computer-based tasks under an assessment engineering paradigm. In D.J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved May 30, 2011 from www.psych.umn.edu/psylabs/CATCentral/.

Martin, J.D. & VanLehn, K. (1993). OLEA: Progress toward a multi-activity, Bayesian student modeler. In S.P. Brna, S. Ohlsson & H. Pain (Eds.), *Artificial intelligence in education: Proceedings of AI-ED 93* (pp. 410–417). Charlottesville, VA: Association for the Advancement of Computing in Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439–483.

Mislevy, R.J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43–71). Hillsdale, NJ: Erlbaum.

Mislevy, R.J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine, 178,* 107-114.

Mislevy, R.J., Behrens, J.T., Bennett, R.E., Demark, S.F., Frezzo, D.C., Levy, R., Robinson, D.H., Rutstein, D.W., Shute, V.J., Stanley, K. & Winters, F.I. (2010). On the roles of external knowledge representations in assessment design. *Journal of Technology, Learning, and Assessment, 8*(2). Retrieved May 31, 2011 from http://escholarship.bc.edu/jtla/vol8/2.

Mislevy, R.J., Behrens, J.T., DiCerbo, K.E. & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and data mining. *Journal of Educational Data Mining, 4,* 11–48.

Mislevy, R.J., Corrigan, S., Oranje, A., DiCerbo, K., John, M., Bauer, M.I., Hoffman, E., von Davier, A.A., Hao, J. (2014). *Psychometric considerations in game-based assessment.* New York: Institute of Play.

Mislevy, R.J. & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Mediated and User-Adapted Interaction, 5*, 253–282.

Mislevy, R.J. & Levy, R. (2007). Bayesian psychometric modeling from an evidence-centered design perspective. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Volume 26* (pp. 839–865). North-Holland: Elsevier.

Mislevy, R.J., Riconscente, M.M., Rutstein, D.W. (2009). *Design patterns for assessing model-based reasoning (Large-Scale Assessment Technical Report 6).* Menlo Park, CA: SRI International. Downloaded May 31, 2011, from http://ecd.sri.com/downloads/ECD_TR6_Model-Based_Reasoning.pdf.

Mislevy, R.J., Steinberg, L.S., Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Johnson, L., Almond, R.A. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*, 363–378.

Neapolitan, R.E. (1990). *Probabilistic reasoning in expert systems: Theory and algorithms.* New York: Wiley.

Ohlsson, S. (1987). Some principles of intelligent tutoring. In R.W. Lawler & M. Yazdani (Eds.), *Artificial intelligence and education* (Vol. 1, pp. 203–237). Norwood, NJ: Ablex.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Kaufmann.

Pennington, N. & Hastie, R. (1991). A cognitive theory of juror decision making: The story model. *Cardozo Law Review, 13*, 519–557.

Schum, D.A. (1994). *The evidential foundations of probabilistic reasoning.* New York: Wiley.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton: Princeton University Press.

Shafer, G. (1987). Probability judgment in artificial intelligence and expert systems. *Statistical Science, 2*, 3–16.
Shafer, G. (1988). The construction of probability arguments. In P. Tillers & E.D. Green (eds.), *Probability and inference in the law of evidence* (pp. 185–204). Dordrecht, The Netherlands.
Shute, V.J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J.D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age Publishers.
Spiegelhalter, D.J. & Cowell, R.G. (1991). Learning in probabilistic expert systems. In J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith (Eds.), *Bayesian Statistics 4* (pp. 447–465). Oxford, U.K.: Oxford University Press.
Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L. & Cowell, R.G. (1993). Bayesian analysis in expert systems. *Statistical Science, 8*, 219–283.
Vendlinski, T.P., Baker, E.L. & Niemi, D. (2008). Templates and objects in authoring problem solving assessments. In E.L. Baker, J. Dickieson, W. Wulfeck & H.F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 309–333). New York: Erlbaum.
Vygotsky, L.S. (1978). *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
Williamson, D. M., Almond, R. G. & Mislevy, R. J. (2000). Model criticism of Bayesian networks with latent variables. In C. Boutilier and M. Goldszmidt (Eds.), *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, 634-643. San Francisco: Morgan Kaufmann.
Woolf, B.P. (2009). Building intelligent interactive tutors. San Francisco: Morgan Kaufmann.

CHAPTER 11 – Reusing Evidence in Assessment and Intelligent Tutors

Diego Zapata-Rivera[1], Keith Brawner[2], G. Tanner. Jackson[1], and Irvin R. Katz[1]
Educational Testing Service[1], US Army Research Laboratory[2]

Introduction

Assessment design methodologies such as evidence-centered design (ECD) provide an approach for representing the "argument" underlying an assessment. Argument-based structures articulate the chain of reasoning connecting task-level data to the evidence required to support assessment claims about the student. Intelligent tutoring systems (ITSs) reflect similar underlying rationale, with decisions about how to update the student model and what to present to the student next, dependent on student performance (evidence). Although the concept of "evidence" is widely used within the assessment community, it is less well articulated in the ITS literature. In this chapter, we argue for the necessity of formalizing the concept of evidence and for implementing tools and services that facilitate the use of evidence in ITSs. Explicitly representing evidence may provide a stronger theoretical basis for an ITS and facilitate the reuse of assessment components across ITSs. Furthermore, additions to the Generalized Intelligent Framework for Tutoring (GIFT) may support defining links from observable evidence to claims (number of links, types of evidence, types of observables, and properties) as well as mechanisms for quantifying evidence. Several examples illustrate how this approach facilitates reuse and handling of evidence in the areas of conversation-based assessment and GIFT-medical prototypes.

Introduction

Understanding how student data are used as evidence to support inferences about students' knowledge, skills, and other attributes (KSAs) is a key aspect in both ITSs and assessment systems. Assessment design usually begins by determining the construct(s) (or KSAs) that are intended to be measured, the claims, and purposes of the assessment; identifying the types of behaviors that we want to observe from students in relation to the target construct(s) (evidence); and defining the types of situations (tasks) through which students will show what they know or can do. When designing these situations/tasks, we need to think about operational constraints (e.g., time and costs associated with production and scoring). Evidence may come from a variety of sources and granularity levels. Following a principled approach to managing, weighing, and quantifying evidence becomes handy when implementing educational assessment systems.

Central to assessment design is an evidentiary argument connecting what students do to different levels of the target constructs (Messick, 1994). Creating these evidence-based arguments is not a trivial task. Fortunately, several assessment design methodologies have been developed to help designers with the creation of evidential arguments underlying assessments. Some of these methodologies include ECD (Mislevy, Steinberg & Almond, 2003), cognitive design system (CDS; Embretson, 1998), and assessment engineering (AE; Luecht, 2013).

ECD provides a principled approach for designing assessments based on the principles of evidentiary reasoning. ECD can be used to design different types of assessments including standardized tests, simulation-based assessments, and performance-based assessments. The main processes of ECD resemble those processes typically used in ITSs (Almond, Steinberg & Mislevy, 2002; Shute & Zapata-Rivera, 2010): activity/task selection based on the student model, presentation of task(s) to the user, response processing (extracting evidence from the student's response and providing task-level feedback), and summary scoring (propagating evidence through the student model and providing summary feedback).

In ECD different aspects of the assessment are conceived as models. ECD includes four different types of models, which correspond to particular assessment questions: the student model (what is being measured?), the evidence model (how is it measured?), the task model (where is it measured?), and the presentation model (how does it look?). These conceptual processes and models provide a common vocabulary to design a variety of assessment applications. Among these ECD models, the evidence model is the one that encapsulates an important part of the evidentiary argument. The evidence model describes how the student model variables should be updated based on the responses to the tasks provided by students (observables). The evidence model includes two components: evidence rules that describe how observables are to be summarized and used as evidence of students performance (task identification and summary of evidence) and the measurement model, which connects the observables to the student model variables (accumulation and synthesis of evidence across tasks). The measurement model is usually implemented using psychometric statistical models.

Formalizing the Concept of Evidence in Its

Processes for eliciting, integrating, maintaining, and making decisions based on evidence are usually distributed among the components of a traditional ITS architecture. For example, the environment/task may collect student actions/responses and send them to the student model, which will use this information to keep an accurate representation of the student's knowledge and skills. The pedagogical model can use information from the environment, the domain model, and the student model to determine how to intervene. Given this distribution of processes and data, explaining why decisions are made and their supporting evidence (e.g., determining the series of actions/responses that were used to support a particular tutoring action) can be difficult.

A clear representation of how particular observables comprise the evidence needed to update the student model can facilitate the development of assessments and ITS. For example, when the pieces of evidence required to support particular claims (e.g., assigning a particular level to any of the student model variables) are identified, assessment or ITS developers can focus on developing the tasks that will allow the students to produce the required pieces of evidence. Also, when making changes to the structure of the assessment, it is useful to identify the claims and corresponding pieces of evidence that need to be modified to end up with a coherent argument structure that supports the new claims. As a mechanism of transparency, it is also important to know how each task is contributing with evidence and how that evidence is used to support particular claims. In terms of student models, this type of functionality could support the development of open student models that in turn could contribute to student learning (Bull & Kay, 2007).

Reusing evidence model information across ITSs is one of the benefits of formally representing the concept of evidence in ITSs. Components of the evidence model could be repurposed when the need for collecting the same type of evidence arises in a different ITS. In addition to having access to the student model information, ITS modules could have direct access to evidence model information and associated tasks, which will support the reuse of these components while maintaining the integrity of the system.

The Case of Conversation-Based Assessment (CBA)

CBA involve students interacting with one or more virtual characters using natural language (i.e., speech or written responses) or menu-based predefined responses. Conversations are designed to assess particular KSAs (constructs). CBAs are intended to be used as a mechanism to gather evidence that may be difficult to obtain using traditional assessment approaches, providing test-takers with multiple opportunities to demonstrate their knowledge/skills, and eliciting explanations about decisions that students make (e.g.,

choices made in a simulation scenario) (Jackson & Zapata-Rivera, 2015; Zapata-Rivera, Jackson & Katz, 2015).

CBAs build on advances in areas such as dialogue-based systems (Adamson, Dyke, Jang & Rosé, 2014; Graesser, Person & Harter, 2001; Millis et al. 2011), human interaction with virtual characters/agents (Chin et al., 2010; Johnson & Lester, 2016), and technology enhanced assessments (Bennett, Persky, Weiss & Jenkins, 2007; Clarke-Midura, Code, Dede, Mayrath & Zap, 2011; Quellmalz et al., 2011).

Conversations in CBAs are designed following the principles of ECD. The intended construct is defined, pieces of evidence required to support student claims are identified, and conversation elements that will provide students with the opportunities to demonstrate their knowledge are designed. Elements involved in the design of conversations include types of virtual characters, questions that will be asked, and character reactions to different types of responses. The resulting product is a conversation space that shows the conversation "paths" that students experience when interacting with the CBA system. Several CBA prototypes have been developed and used to measure skills such as science inquiry and other constructs (Jackson & Zapata-Rivera, 2015).

Case Study: Reuse of Evidence in the Volcano and Weather Prototypes

Two isomorphic CBA prototypes (the Volcano and the Weather) were developed and used to explore reusability issues. In this section, we describe these prototypes and how evidence and other components of the first system were repurposed in the creation of the second one.

The Volcano Prototype

The Volcano prototype is a CBA designed to measure science inquiry skills in the context of collecting data to predict a volcano eruption (Figure 1). Virtual characters provide information about volcanic eruption, the use of seismometers to collect data, and the criteria for establishing eruption alert levels. After learning about volcanoes, students engage in data collection by placing seismometers on a simulated volcano. Students can choose up to four seismometers. Students also determine the data collection time. Virtual characters provide feedback and ask questions about the decisions students make (e.g., "Why did you select X seismometers?" "Which note would you keep to make predictions?" "Do you agree with Art's prediction? Why?"). Depending on the student response, the virtual character responds by asking follow-up questions or providing additional information (e.g., rephrasing the question, asking for additional information, or clarifications) (Zapata-Rivera et al., 2016).

Figure 3. A screenshot of the Volcano CBA prototype. Lucas is asked which note he would keep to support his predictions later.

Evidence in the Volcano Prototype

Some of the pieces of evidence used in the Volcano CBA include the following:

- quality of description of the process of volcano eruption and volcanic formation (earth science knowledge)
- accuracy of selected sequence of volcanic seismic events (analyzing data and identifying patterns)
- accuracy of identification and quality of description of high-frequency and low-frequency seismic events based on seismometer data (analyzing data and identifying patterns)
- quality of selection between two data collection notes supported by different amounts of data (conducting data collection)
- accuracy of prediction based on data collected (making predictions based on data)
- quality of explanation and demonstration of how data collected are used to support a volcanic prediction (making predictions based on data)

Assessment design patterns can be used to describe and document the evidentiary argument in a narrative form (Mislevy & Haertel, 2006). Design patterns describe the KSAs that are assessed, the features of the situations that are used to elicit required evidence, and the types of behaviors that we expect to observe from students and will count as evidence for the intended KSAs. The design pattern of the Volcano CBA

focused on the skills of planning and carrying out data collection in a virtual field and using collected data as evidence to predict a natural event. The Volcano design pattern includes a subset of components of scientific reasoning (using observation data to make a prediction for natural events) and includes four focal KSAs: earth science knowledge, analyzing data and identifying patterns, conducting data collection, and making predictions based on data. Conversations and other items in the volcano scenario provide the evidence necessary to make claims about students' levels on each of these four KSAs (Liu, Steinberg, Qureshi, Bejar & Yan, 2016; Zapata-Rivera et al., 2016).

By looking at these design patterns for the Volcano prototype one can see how the observables are used to support particular assessment claims as well as to identify assessment claims that lack enough supporting evidence. In addition, enhanced score reports can be produced to provide additional information regarding the evidence used to produce the scores.

Leveraging this work and evidence identification within the volcano activity, a parallel task was developed using the same design pattern in the area of weather exploration (collecting data and making predictions of the likelihood of a thunderstorm).

Reusing Evidence in the Weather Prototype

The Weather CBA prototype was developed with the goal of exploring the feasibility of reusing different aspects of the Volcano CBA prototype (e.g., assessment documentation and graphical components) to create a new isomorphic task and evaluating how students would perform on each of these prototypes. The Weather CBA prototype is intended to measure the same KSAs as the Volcano prototype.

The design of the Weather CBA prototype was informed by following the assessment design documentation available for the Volcano CBA scenario (e.g., design pattern information) to change the surface level features while retaining the same underlying evidence structure. Thus, both CBA prototypes share the same types of tasks, structure and conversations. However, content aspects needed to be developed to accommodate them for the new context. Figure 2 shows a screenshot of the Weather CBA prototype, which shows a conversation between the student and virtual characters about the quality of data collection notes similar to the one shown in Figure 1. In this case, the conversation refers to weather stations rather than seismometers.

Figure 2. A screenshot of the Weather CBA prototype. Lucas is asked which note he would keep to support his predictions later.

Some of the pieces of evidence used in the Weather CBA include the following:

- quality of description of the elements of a thunderstorm and its formation (earth science knowledge)
- accuracy of selected sequence of events leading to a thunderstorm (analyzing data and identifying patterns)
- accuracy of identification and quality of description of levels of water vapor and instability and the presence of cold fronts based on weather station data (analyzing data and identifying patterns)
- quality of selection between two data collection notes supported by different amounts of data (conducting data collection)
- accuracy of prediction based on data collected (making predictions based on data)
- quality of explanation and demonstration of how data collected are used to support a thunderstorm prediction (making predictions based on data)

A study comparing the psychometric properties of these two CBA prototypes (N = 210 students) showed evidence of the comparability of these tasks based on results from item analyses and factor analyses. Results showed similar discrimination and difficulty levels of items across the two prototypes. Results of confirmatory factor analysis suggests the presence of four constructs. The results also showed potential for using CBAs to measure science inquiry skills across different contexts (Liu et al., 2016).

The development of the second CBA resulted in significant cost savings due to the amount of reuse, which included the use of existing graphical components, code, measurement instruments, and documentation.

Case Study: Reuse of Evidence in the Gift-Medical Prototypes

The Generalized Intelligent Framework for Tutoring (GIFT) domain module contains all of the domain information in accordance with the original design principles (Sottilare, Brawner, Goldberg & Holden, 2012; Sottilare, Brawner, Sinatra & Johnston, 2017). This module encapsulates domain-specific information, especially assessment-specific information. The domain module is configured by the domain knowledge file (DKF), which has an extensible markup language (XML)-based file structure. As a byproduct of this design decision, the following primary information that it has, examples of each, and concrete examples of content for a medical task occur:

- A list of concepts and subconcepts to be instructed
    - Example: “stay with unit” concept
    - Example: “move casualties to a safe area” concept
- Content to present on the concepts
    - HTML pages, pictures, videos
        - Example: a PowerPoint Show (.pps) file for how to stay with the unit
        - Example: a video on the importance of moving casualties
    - Variety of scenarios, terrains, situations
        - Example: a scenario which involves the concept of moving with the unit and moving casualties (this scenario teaches multiple concepts)
- Assessment for the content
    - Quiz, test
        - Example: 3 test questions to gauge understanding of moving with the unit and moving casualties
    - Constraints, assessment rules
        - Example: logic, within a scenario, which assesses performance staying with a unit (player.location must be ≤ 20 meters from calculated unit centroid)
        - Example: logic, within a scenario, which assesses performance of whether an injured unit is within a dangerous location for longer than 30 seconds (injuredPlayer.location must be ≤ 30 seconds in location.unsafe)
- Feedback for the content
    - Hints, remedial material

        - Example: replaying the .pps of staying with the unit
        - Example: a picture of what happens to people to fail to move casualties
    - Hints, prompts, scenario changes
        - Example: “You are part of a unit, you need to stay close to them.”
        - Example: “Move the injured soldiers to a safer location.”

The configuration and specification of the DKF lends itself to relatively easy reuse of domain-specific content with minimal changes to the underlying structure. Within a simulation, the DKF schema naturally lends itself toward an ECD approach. When actions are observed which meet the conditions set out in condition classes, learner performance is noted, modeled, and communicated onwards to the learner module. In this manner, assertion of individual concept performance is only communicated when there is evidence of its demonstration. As an example, in the Tactical Combat Casualty Care (TC3Sim; Sotomayor, 2010), a military medical simulation that implements the GIFT model, there are specific domain knowledge constraints that apply to many scenarios. Examples of these include the following:

- Casualties should be safe before they are treated.
- The area directly above a missing limb should receive a tourniquet to prevent blood loss.

The domain-specific, evidence-based constraints created for TC3Sim, and their coupled feedback, are used across a large number of scenarios because they are relevant in a number of disparate situations. The assessment logic is reused significantly across the simulator, while still being decoupled from the simulator itself, depending on how it was authored. The same constraints were used in the context of another simulator, Virtual BattleSpace 2 (VBS2; Bohemia Interactive Australia, 2012), without a change in the authoring tool or underlying logic (US Army Research Laboratory, 2017).

Other Military Use Cases

Generally speaking, many modern military simulators use a shared protocol, with the majority choosing to use either the Distributed Interactive Simulation (DIS; IEEE 1278.1-2012, 2012) or high-level architecture (HLA; SISO-STD-004.1-2004, 2004) standards. This makes the link between the ITS and the training environment significantly reusable. Furthermore, because of the standardization of the simulator-tutor link, the assessment rules used in any domain can be reused to a new domain to add complexity. Consider a simple example scenario that involves investigating a residence where a scenario-specific constraint such as “don’t shoot anyone” is in effect. The abovementioned medical constraints can be layered on top of the existing scenario with minimal work. The separation of the assessment logic from the simulator and the student model allows for significant reuse. The reuse across similar interfaces allows for many different portable assessments, as DIS/HLA is used across a wide variety of simulators, including live operations, virtual operations, and constructive simulations in undersea warfare, flight simulators, and ground operations. In these examples, although the same assessment component is used, the appropriate observable behavior needs to be identified and mapped to the specific situation. For example, not shooting anyone is a universally understood behavior and the same actions (i.e., pulling the trigger) apply regardless of context or setting. However, with a more nuanced behavior like removing injured personnel from an active battle situation before administering aid could be exhibited through very different behaviors. For example, getting someone to safety in an urban context might mean getting into a room and closing a door, whereas in a more rural context it might mean removing line of sight by moving behind a large obstacle (e.g., geographic barrier). These behaviors accomplish the same assessment goal of getting injured personnel to safety before

applying aid, but the specific pieces of evidence within the environment need to be explicitly tied to that assessment goal since they require different context specific actions. Mapping these context specific actions to the construct(s) of interest is the primary purpose of an evidence layer and facilitates reuse of concepts and evidence across situations through clear and explicit links.

Implications for GIFT: Evidence Model Layer

ECD and other assessment design framework methodologies show that having a representation of how observed data are used to support assessment claims facilitates the creation of assessments that implement a sound argument structure. In addition to facilitating reuse of evidence, additional benefits of an evidence model layer for ITS include the following:

- *Integrating claims and evidence in reporting information produced by the tutor*. Assessment claims and supporting evidence could be shared with different types of users to support particular ITS processes. For example, reports can be produced that show evidence available with the current tasks and the types of ITS decisions that are supported with these set of tasks. Other reports inform the development of new content and associated tasks, share information of the student model (assessment claims), and share pieces of evidence collected by the system with teachers and students.

- *Automatic evaluation of evidence*. If an evidence model layer is followed, it is possible to create tools that can perform multiple functions. These functions include the identification of evidence criteria, identification of missing pieces of evidence, and the ability to quantify the amount of evidence available. Each of these functions can then support the claims or decisions made by the tutoring system. These tools can take into account the number and type of links from observables to claims available, the types of observables and other properties that characterize the pieces of evidence collected (e.g., information about the context/tasks used to collect the evidence).

The next section describes how the features of a separate evidence model could be integrated into the GIFT architecture.

Exploring the Implementation of an Evidence Model in GIFT

The GIFT architecture is a flexible open architecture that could support a formal representation of an evidence model or a similar module. The design of the GIFT system has several overarching architectural goals. These include 1) the ability to rapidly transition new models from research to use; 2) the adjustment of assessment logic without reengineering; 3) the ability to easily add complexity, content, models, or new tasks; and 4) the need to be open source and freely available (Brawner, Sinatra & Sottilare, in press). Fundamentally, accomplishing these goals relies upon the availability of an open-source system of interchangeable parts.

In GIFT, it is possible to encapsulate each of the models of a traditional ITS architecture within a standardized software process to allow for their easy interchange. By encapsulating some of the aspects that deal with evidence in the GIFT framework, it is possible to implement a component similar to the evidence model. This may be the use, handling, and sharing of evidence in GIFT.

Aspects of the GIFT domain module, the learner module, and the pedagogical module can be integrated into a module that plays a role similar to that of the evidence model in ECD-based assessments. In fact, current work on The Training Learning Architecture (TLA) project has started looking at these types of synergies (Regan, 2013). The TLA project has grown out of the increasing need to be able to share learner information and experiences across systems. At its heart is the experience API (xAPI), which operates on

the actor-verb-object description of learner experiences, and shares its information with a learner record store (LRS). As an example, a statement might be issued that "Learner1" "mastered" "physics", encoded as appropriate to the standard. The xAPI forms the backbone of sharing learner information across systems in the same manner that the Sharable Content Object Reference Model (SCORM; ADL, 2001) standard enabled the sharing of content across systems.

As part of the growing TLA project, there are emerging APIs to reason about the xAPI information stored within a LRS. Among these are the performance API (pAPI), and evidence mapper API (eAPI), which aim to be able to share learner profiles across systems, and update a learner profiles based on evidence, respectively. The objective is for both of these systems to have the underlying reasoning algorithms based on the ECD framework to update and share experiences across many different systems.

Other efforts of incorporating domain-independent feedback and performance annotation are being examined by the US Air Force Research Laboratory (Galster & Johnson, 2013). This project has the goal of developing assessment information from simulator performance and physical sensors in a manner agnostic to particular simulators. It has experienced a modicum of success and is seeking early standardization.

Conclusions

The examples presented in this chapter illustrate how evidence can be reused to create additional tasks or support evaluation and development processes. The use of ECD principles in the form of assessment design patterns in CBA facilitates the identification and reuse of evidence. Current work on the TLA project explores similar ideas.

A separate evidence model or a similar model in the GIFT architecture can be used to align content and evidence needs. This can be done by providing information about the content, feedback, and tasks required to cover the previously defined evidence. Furthermore, it supports the evaluation of the system by keeping an up-to-date list of KSAs covered and their status (e.g., for which KSAs the system includes enough high quality evidence to allow for more robust, valid, and supportable claims). Also, it can facilitate reuse of various aspects of the model including content, feedback, and tasks associated with particular pieces of evidence.

Given the flexibility of the GIFT architecture, it may be possible to encapsulate various ITS components to improve the current support for handling and sharing evidence. This may result in enhanced reporting mechanisms available for developers and users (e.g., by including additional information on the reports such as the evidence used by the tutor to make particular decisions), and a faster development and maintenance cycle (e.g., by identifying assessment claims that lack sufficient evidence).

References

Adamson, D., Dyke, G., Jang, H.J. & Rosé, C.P. (2014). Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of AI in Education*, 24(1):91–121.

Advanced Distributed Learning (ADL). (2001). Sharable Content Object Reference Model (SCORM™). Advanced Distributed Learning. Retrieved from http://www.adlnet.org.

Almond, R.G., Steinberg, L.S. & Mislevy, R.J. (2002). A four-process architecture for assessment delivery, with connections to assessment design. *Journal of Technology, Learning, and Assessment*, 1(5), 1–64.

US Army Research Laboratory. (2017). GIFT Domain Knowledge File. Retrieved from https://gifttutoring.org/projects/gift/wiki/Domain_Knowledge_File_2017-1.

Bennett, R. E., Persky, H., Weiss, A. & Jenkins, F. (2007). *Problem-Solving in technology rich environments: A report from the NAEP technology-based assessment project*. NCES 2007-466, US Department of Education, National Center for Educational Statistics, US Government Printing Office, Washington, DC.

Bohemia Interactive Australia. (2012). White paper: VBS2 (release version 2.0). Retrieved from https://bisimulations.com/VBS2_Whitepaper.pdf.

Brawner, K., Sinatra, A. & Sottilare, (in press). Motivation and Research. In Architectural Intelligent Tutoring, *International Journal of Simulation and Process Modeling*, 2016 in press.

Bull, S., and Kay, J. (2007). Student models that invite the learner in: the SMILI open learner modelling framework. *International Journal of Artificial Intelligence in Education*, 17(2), 89–120.

Chin, D.B., Dohmen, I.M., Cheng, B.H., Oppezzo, M., Chase, C. & Schwartz, D. (2010) Preparing students for future learning with Teachable Agents. *Education Technology Research and Development*, 58(6): 649–669.

Clarke-Midura, J., Code, J., Dede, C., Mayrath, M. & Zap, N. (2011). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In M.C. Mayrath, J. Clarke-Midura & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age. 125–47.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 300–396. doi:10.1037/1082-989X.3.3.380

Galster, S. M., and Johnson, E. M. (2013). Sense-assess-augment: A taxonomy for human effectiveness (No. AFRL-RH-WP-TM-2013-0002). Air Force Research Lab, Wright-Patterson AFB, OH, Human Effectiveness Directorate.

Graesser, A. C., Person, N. K. & Harter, D. (2001). The Tutoring Research Group: Teaching tactics and dialogue in AutoTutor. *International Journal of Artificial Intelligence in Education*. 12, 257–279.

IEEE 1278.1-2012, (2012). IEEE Standard for Distributed Interactive Simulation--Application Protocols. Retrieved from https://standards.ieee.org/findstds/standard/1278.1-2012.html.

Jackson, T., and Zapata-Rivera, D. (2015). Conversation-based Assessment. *R&D Connections. No 25*. Princeton, NJ: ETS.

Johnson, W.L., and Lester, J.C. (2016). Face-to-Face Interaction with Pedagogical Agents, Twenty Years Later. *International Journal of Artificial Intelligence in Education*. 26 (1), 25–36.

Liu, L., Steinberg, J., Qureshi, F., Bejar, I. & Yan, F. (2016). Conversation-based Assessments: An Innovative Approach to Measure Scientific Reasoning. *Bulletin of the IEEE Technical Committee on Learning Technology*. 18(1), 10–13.

Luecht, R. M. (2013). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 59–76). New York, NY: Routledge.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.

Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. C. & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou & J. Lakhmi (Eds.), *Serious games and edutainment applications* (pp.169–196). London, UK: Springer-Verlag.

Mislevy, R.J., and Haertel, G.D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, 25, 6–20.

Mislevy, R.J., Steinberg, L.S. & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*. 1, 3–62.

Quellmalz, E.S., Timms, M.J., Buckley, B.C., Davenport, J., Loveland, M. & Silberglitt, M. D. (2011). 21st Century Dynamic Assessment. In M.C. Mayrath, J. Clarke-Midura & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age. 55–90.

Regan, D. A. (2013). The Training and Learning Architecture: Infrastructure for the Future of Learning. Paper presented at the Invited Keynote International Symposium on Information Technology and Communication in Education (SINTICE), Madrid, Spain.

Shute, V. J., and Zapata-Rivera, D. (2010). Intelligent Systems. In E. Baker, B. McGaw & P. Peterson (Eds.), *Third Edition of the International Encyclopedia of Education*. Oxford, UK: Elsevier Publishers. vol. 4, pp. 75–80.

SISO-STD-004.1-2004. (2004). Standard for Dynamic Link Compatible HLA API Standard for the HLA Interface Specification (IEEE 1516.1 Version) (reaffirmed 8 Dec 2014).

Sotomayor, T.M. (2010). Teaching tactical combat casualty care using the TC3 sim game-based simulation: a study to measure training effectiveness. *Studies in Health Technology and Informatics*, 154, 176–9.

Sottilare, R.A., Brawner, K.W., Goldberg, B.S. & Holden, H.K. (2012). The generalized intelligent framework for tutoring (GIFT). Orlando, FL: US Army Research Laboratory, Human Research & Engineering Directorate (ARL-HRED).

Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017. DOI: 10.13140/RG.2.2.12941.54244.

Zapata-Rivera, D., Jackson, T. & Katz, I.R. (2015). Authoring Conversation-based Assessment Scenarios. In R. A. Sottilare, A. C. Graesser, X. Hu, and K. Brawner (Eds.), *Design Recommendations for Intelligent Tutoring Systems Volume 3: Authoring Tools and Expert Modeling Techniques*. (pp. 169–178). US Army Research Laboratory.

Zapata-Rivera, D., Liu, L., Chen, L., Hao, J. & von Davier, A. (2016). Assessing Science Inquiry Skills in Immersive, Conversation-based Systems. In B. K. Daniel (ed.), *Big Data and Learning Analytics in Higher Education*, Springer International Publishing, 237-252, doi:10.1007/978-3-319-06520-5_14.

CHAPTER 12 – Methods for Assessing Inquiry: Machine-learned and Theoretical

Michelle LaMar[1], Ryan S. Baker[2], and Samuel Greiff[3]
Educational Testing Service[1], University of Pennsylvania[2], University of Luxembourg[3]

Introduction

Inquiry skills are critical to virtually any kind of problem solving and particularly to the practice of science. However, these skills cannot be easily assessed using traditional static assessment items, in which it is difficult to follow (and capture) the actual inquiry process. Given a sufficiently interactive task, the challenge becomes how to identify and score the inquiry skills applied to the task. This chapter examines three different methods for assessing inquiry skills using the process data generated by computerized interactive tasks. The first two methods identify inquiry strategy-use based on specific features in the data. The first uses theoretically defined features, whereas the second leverages machine learning to identify and combine relevant features. The third method uses generative process models to compare student actions to probabilistic agents implementing targeted inquiry strategies. The advantages and drawbacks of each method are discussed along with assessment contexts that favor particular approaches. Implications for detection and scoring of inquiry within the Generalized Intelligent Framework for Tutoring (GIFT) are addressed.

In a world that increasingly uses technology in the workplace, at home, and as a medium for commerce and communication, the ability to think scientifically and solve technical problems is increasingly useful in modern life. Both scientific inquiry processes and problem-solving skills have been identified as key 21st century skills due to their pivotal role for success in contemporary societies (Dede, 2010). As with many 21st century skills, these competencies are inherently interactive, involving multi-step processes that must adapt to information gathered and results generated. In K–12 education, the need to teach authentic science practices has been a major impetus for the move from static textbooks and cookbook labs to more interactive science instruction (e.g., Next Generation Science Standards [NGSS; NGSS Lead States, 2013]) that frequently use intelligent tutoring systems (ITSs), computer simulations, and games. With this shift in instructional emphasis comes a need to also assess these skills using interactive problem-solving tasks. ITSs in particular depend upon real-time skill diagnosis and assessment to allow for customized instruction and guidance.

Scoring interactive performances, however, is much more complicated than scoring responses chosen from a small, fixed list of options such as in the multiple-choice item formats favored by standardized testing. While performance assessment is not new, evaluating performance has traditionally been conducted by human raters who call upon complex experience to interpret and judge the amount of skill and understanding displayed in a particular performance. Machine scoring has made major progress, mostly in the form of scoring written text (Shermis, 2014; Liu et al., 2016), but more recently there has been progress in scoring competencies of complex problem solving or science inquiry with interactive tasks (Gobert, Sao Pedro, Baker, Toto & Montalvo, 2012; Greiff, Wüstenberg & Avvisati, 2015). In this chapter, we discuss methods for automatically identifying and scoring inquiry strategies based on logged interactions within computerized simulations.

Science Inquiry

Science inquiry and problem solving have frequently been characterized in terms of two major phases: hypothesis generation and hypothesis testing (Popper, 1959; Klahr, 2002). While many aspects of practicing

science could be demonstrated using interactive science tasks, the focus of this chapter is on inquiry strategies that might be used in the hypothesis-testing phase of scientific investigations. Successful strategies in scientific hypothesis testing largely overlap with strategies that are useful for complex problem solving (Jonassen, 2007; Klahr & Dunbar, 1988), thus making the detection and assessment of these inquiry strategies applicable well beyond traditional science instruction.

The definition of science practices in NGSS (NGSS Lead States, 2013) includes eight science practices, of which hypothesis testing is mostly covered by two practices: "planning and carrying out investigations" and "analyzing and interpreting data". Key skills listed in these practices include identifying, controlling, and measuring relevant variables, deciding what and how much data to collect under what range of conditions, conducting systematic analysis of data, recognizing evidence that contradicts the hypothesis, and evaluating the strength of conclusions that can be made from a set of data. Different theoretical frameworks break these skills down in different ways with different emphases (for example, see Wieman, 2015). However, the key elements of investigation remain: how to collect data and how to interpret data.

Assessment and Scoring Challenges

To assess skills in science inquiry and problem solving, it is necessary to have the examinees engage in the process of science inquiry or problem solving. The design of appropriately interactive tasks to allow for the demonstration of these skills comprises the first assessment challenge. Traditional test item design favors testing one skill at a time; however, the interdependency between the inquiry practices and the time commitment required for an interactive task can make it more feasible to assess multiple skills simultaneously within a larger investigation. The impulse to limit the task, ensuring that we understand what we are measuring, pushes design toward more scaffolded tasks with a fixed number of choices with limited options. The desire to measure more authentic inquiry including interdependent skills and content knowledge, on the other hand, pushes the design to open, exploratory, scenarios in which there are many choices and many ways to go wrong. Depending upon the purpose of the assessment, a more scaffolded versus a more open design are considered but in practice the choice is often made based on the limitations of the scoring.

A second assessment challenge is scoring performances in inquiry investigation. The easiest scoring method is based on the outcome of the investigation alone. If the examinee is able to draw the correct conclusions, we can infer that they probably carried out a successful investigation. There are a number of problems with the outcome-only metric, however. While correct answers undoubtedly correlate with correct practices, it is also possible to stumble upon the correct answers by chance or, in more scaffolded tasks, to try every possible variable combination without taking the time to develop a systematic data collection plan. Furthermore, an incorrect answer might stem from a failure of any number of skills or content knowledge, especially in the more open-ended task designs. Thus, an incorrect answer could easily mask good overall inquiry skills.

As an alternative to outcome-only scoring, a record of the actual steps taken within an investigative task can be analyzed for evidence of science-inquiry skills. This approach is particularly compatible with computerized interactive tasks, which allow for a wide variety of experimental setups along with the collection of data about actions taken within the task. Scoring this collection of process data, however, is far from straightforward. For each decision the student is allowed, the number of paths through the task grows exponentially. Thus, scoring on the basis of a set of "correct" paths quickly becomes impractical. The methods presented here involve analysis of actions taken within an interactive task to detect instances of good science practice and provide scoring data for generalized problem-solving tasks. This approach requires defining not only what constitutes good inquiry, but also what the application of good inquiry skills looks like in specific task contexts. Further, detecting an instance of inquiry strategy use is not, in itself, a score.

Further analysis or a broader model might be needed to connect the identified behavior to valid inferences about the student's problem-solving ability.

In the following sections, we describe three different methods for identifying and scoring good inquiry practice from a record of interactions within a complex task. The first two methods identify inquiry strategy-use based on specific features in the data. The first uses theoretically defined features, whereas the second leverages machine learning to identify and combine relevant features. The third method creates probabilistic models of decision making based on different inquiry strategies and compares student actions to model predictions to identify strategy use.

Methods for Identifying Good Inquiry Strategy

Theoretically Defined Performance Indicators

A first approach toward identifying inquiry skills in computer-simulated task environments is based on theoretically developed behavioral indicators that are directly derived from the underlying definition of science inquiry and its theoretical components. In the theoretical approach, a definition of science inquiry is the starting point (Kuhn, 2012; Zimmerman, 2007). Along with this definition, carefully drafted tasks aimed at measuring science inquiry need to be developed. These tasks need to follow the theoretical rationale and incorporate those aspects of science inquiry that are considered crucial in the definition. Equipped with an adequate task environment, specific behaviors (or behavioral patterns) that can occur within the task environment are identified as indicating either high or low levels of science-inquiry skills. Thus, it is important that the task space is designed in a way that it allows for different behaviors that are indicative of the underlying theoretical conception and that represent high versus low proficiency levels of this conception.

A straightforward example of an important concept in the field of science inquiry is the principle of isolated variation (sometimes referred to as the vary-one-thing-at-a-time strategy [VOTAT]; Tschirgi, 1980) in which students demonstrate their ability to comprehend, use, and argue along the lines of causal relations within scientific phenomena (Kuhn, Black, Keselman & Kaplan, 2000). Because the VOTAT strategy is effective in isolating causal relationships and reducing the influence of confounding variables, it is considered relevant across a number of domains and has been identified as an important strategy in the field of complex problem solving and science inquiry (Wüstenberg, Greiff & Funke, 2012). Interestingly, Kuhn (2012) highlights that even adults often suffer from an insufficient understanding of the principle of isolated variation.

Figure 1 displays an example of a task environment taken from the field of complex problem solving, though it could easily be used as a science-inquiry task with a couple of small adaptations. This type of task, often referred to as MicroDYN-type of task (Greiff, Wüstenberg & Funke, 2012), features a small simulated system with multiple inputs or controls and multiple outputs or effects. The goal set for the student is to map the relationship between the input and output variables. The example in Figure 1 is usually used in a secondary education context and students are asked to determine how different components of a windmill affect both the noise produced by the windmill and the costs associated with operating it. The problem environment presented here is not fully open because students have a small number of actions they can perform in the environment. However, there is no explicit guidance of student approaches, which allows expression of individual differences between students. There are several strategies to solve this task, but students who are familiar with and proficient in applying and using the VOTAT-strategy usually perform better on these tasks. This is no coincidence because the task environment was developed on the basis of a theoretical understanding of complex problem solving and the types of inquiry skills that are needed to get from the initial state to the goal state. The theory, which highlights the principle of isolated variation as an

important aspect of both science inquiry and complex problem solving, ensured the development of a task in which VOTAT behavior would be both productive and distinguishable from non-VOTAT behavior. Scoring for the task would be based on whether the examinees apply the principle of isolated variation during the unguided exploration phase. The final step is to define an indicator of the targeted behavior (here VOTAT) from the information stored in computer-generated log files. A theory-driven approach can be used along with other inquiry principles to do this. The common umbrella would be to develop the task design, targeted behaviors, and behavioral indicators in such a way that they elicit evidence of understanding and competency of the theoretically defined underlying principles.

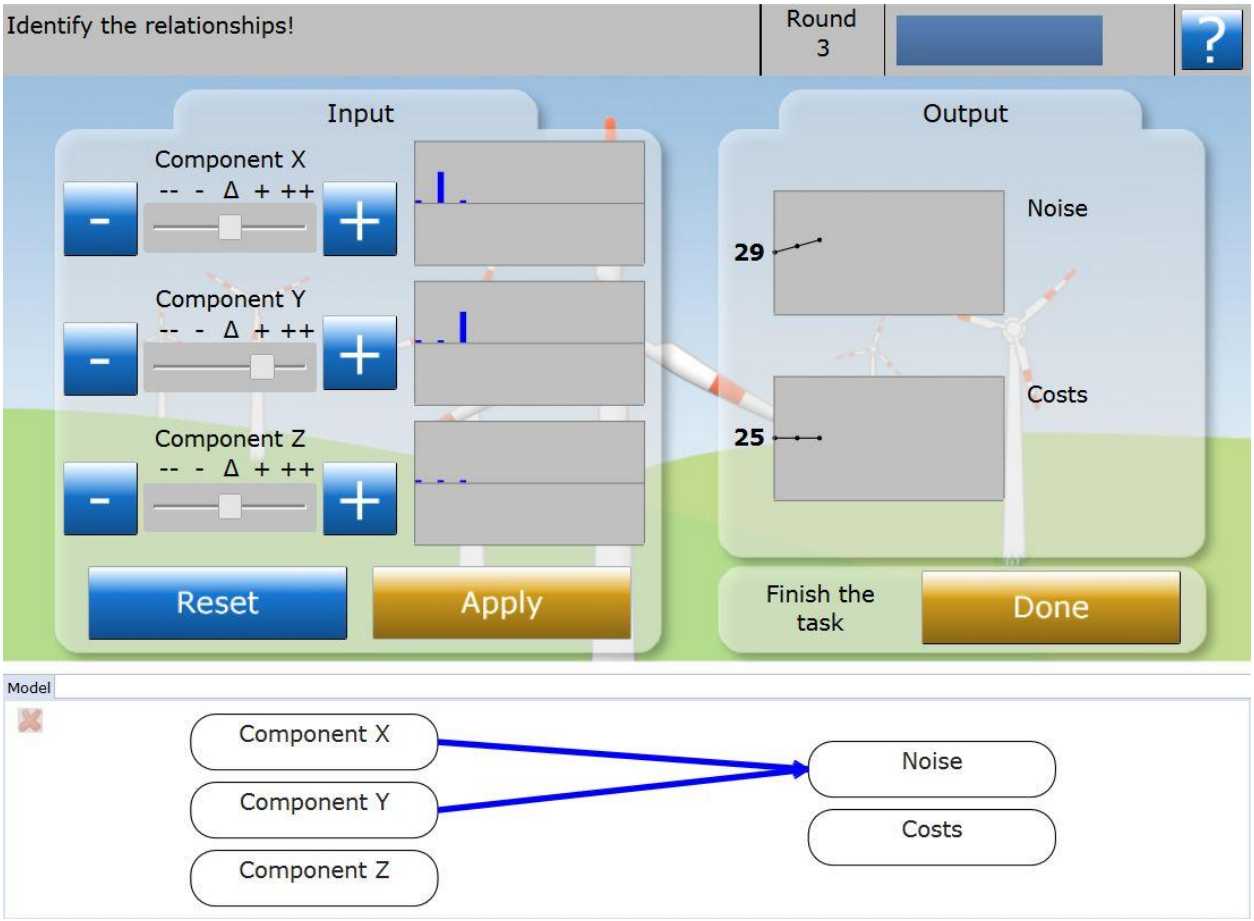


Figure 1. Complex problem-solving environment within the MicroDYN approach. Students are asked to discover the relations between the inputs on the left side and the outputs on the right side during a phase of unguided exploration.

Quantifying Use of the Principle of Isolated Variation (VOTAT-Strategy)

The operationalization of science inquiry is still a matter of some theoretical debate. Even with a profound theoretical understanding, there are many ways isolated variation as an overarching principle might be scored in different (or even the same) task environments. Depending on the specific scoring, results might vary widely. In the following, some empirical examples are presented, but it is important to note that these indicators have been theoretically defined based on the specific assumptions and goals of this assessment and manifest within the scoring practices limited by those assumptions and goals.

Vollmeyer, Burns, and Holyoak (1996) reported on experimental studies in which they investigate the impact of the use of isolated variation in a problem-solving environment called Biology-Lab, which includes four independent and four dependent variables. They differentiate between students who initially use or adopt isolated variation over the course of problem exploration from those who use other, less efficient strategies. Students were allowed four learning rounds, which consisted of six experimental trials per round. A round was coded as using VOTAT if, in at least four out of the six trials, only one input variable was varied while the others were set to zero. The general pattern of results indicated that use of an efficient strategy (here VOTAT) as compared to other strategies is associated with better learning outcomes and with a higher level of rule acquisition as an overall performance indicator. Thus, a theoretically defined process indicator of isolated variation was shown to relate to performance indicators within a problem-solving environment, thereby confirming the initial theoretical assumption.

In a similar vein, Greiff, Wüstenberg, and Avvisati (2015) analyzed log file data from the 2012 assessment of problem solving in one of the most important educational large-scale assessments, the Programme for International Student Assessment (PISA; OECD, 2014). In a large sample of 15-year-old students, Greiff et al. scored whether students employed the principle of isolated variation for all of the three input variables during the unguided exploration phase (i.e., consistent use of VOTAT) or whether they did so either not at all or inconsistently (i.e., not for all of the input variables). They reported that the strategic use of isolated variation in a MicroDYN-like task (see Figure 1) was related to both overall task performance and the overall problem-solving score in PISA. That is, the specific strategic behavior within one task was related not only to performance in this task but beyond that one task to overall problem-solving performance (and thus to general problem-solving and inquiry performance).

In explorative analyses, Greiff et al. (2015) employed an additional and more detailed scoring in which student strategy was scored along the number of input variables for which the principle of isolated variation was applied. Thus, scores ranged from 0 to 3 for the task with 3 input variables. In doing so, the authors (tentatively) reported progressive proficiency levels among students. These proficiency levels went beyond merely distinguishing between those students who used and those who did not use isolated variation as in the first analyses. Of note, since two different ways of scoring were used in the same sample, slightly different insights were gained depending on the specific way the principle of isolated variation was operationalized.

The two studies mentioned here serve as examples that consistently show the relation between the use and application of the principle of isolated variation as an important conceptual aspect of inquiry skills, its operationalization through behavior-based indicators, and external criteria that serve as validity evidence. However, there are both pros and cons of this theoretically motivated approach, which are briefly discussed in the next section.

Pros and Cons

An important advantage of the above-described approach of employing theoretically defined indicators of inquiry skills is the direct connection to theories of the human mind in general and theories on inquiry and problem solving in particular. Because it is necessary to engage in elaboration of the underlying theoretical foundation before any indicator can be clearly defined, all indicators are easily interpretable. In contrast to the machine-learning approaches that are described in the next section, theoretically defined performance indicators are always embedded in some kind of broader framework that helps put the specific indicators into perspective and give them meaning.

The theoretical approach is associated with drawbacks as well. One of them certainly is that in many cases a specific behavior (or, even more so, a non-behavior) cannot be directly mapped to an underlying theoretical defined construct. For instance, it is quite straightforward to claim – against the backdrop of the above-

mentioned empirical studies – that there is some causal connection between the use of VOTAT and performance in problem-solving environments. However, what about students who did not employ this principle? There might be several reasons for their lack of adequate strategic behavior including a lack of understanding, low level of motivation and task engagement, or issues with understanding the instructions of the task. Thus, it is often difficult to establish an isomorphic mapping between a theoretical concept and a specific behavioral indicator. In addition to this, complex strategies that involve a number of variables and indicators are difficult to detect in a purely theoretical approach because it requires a clear understanding of the specific underlying mechanisms and how they manifest as specific behaviors. Thus, the theoretical approach is often somewhat limited to a narrow set of theoretical aspects with the complex interplay of several behaviors being omitted. This last shortcoming is exemplified by the fact that the large body of literature on science inquiry mostly revolves around the rather straightforward and easily to define and detect principle of isolated variation (Kuhn, 2012). Overall, theory is the cornerstone for any sound and scientifically valid understanding. However, this type of confirmatory approach needs to be complemented by more data-driven and exploratory methods such as machine-learned inquiry detection.

Machine-Learned Inquiry Detection

A second approach toward identifying effective inquiry strategies is to use machine learning, often referred to in the domain of education as either educational data mining or learning analytics (Baker & Siemens, 2014). This research area refers to a broad range of methods that leverage the potential of analyzing thousands of possible relationships among variables in an automated fashion. Educational data-mining methods are particularly useful either when there is relatively little known about the domain being analyzed (in which case "unsupervised" methods are used that do not privilege specific variables for analysis) or when the construct to be modeled is known but it is thought that the best prediction or inference of it will involve combinations of variables that are more complex than a human could reasonably identify (in which "supervised" methods are used that attempt to discover the best combination of variables that identifies a known variable). Supervised methods, such as classification, are used to categorize specific cases into one of a small number of known categories on the basis of "features" of the data for that case. Machine-learning approaches have the advantage that the relationship between the features and the categories do not need to be known in advance; the set of features and their interactions are selected by the algorithm based on the relationships found in the data. In this section, we describe two successful uses of supervised methods to identify student inquiry skill. In one of these, humans identified successful inquiry in a limited data set and then machine learning was used to replicate their judgments at scale. In the second, successful inquiry was identified as student success at solving a puzzle that required inquiry and machine learning was used to identify patterns of behavior that led to that successful performance.

Replicating Human Judgment

In this first example, humans identified inquiry within a limited data set, and then machine learning was used to replicate that judgment. This work was predicated on the assumption that expert researchers in inquiry could recognize appropriate inquiry when they saw it, but that transforming that comprehension into simple rules is challenging and may lead to overly precise rules that either exclude some acceptable strategic behavior or treat some inappropriate strategies as appropriate. For instance, the VOTAT rule does not clarify how to treat a case where a student runs an experiment, changes two variables, runs an experiment, changes one back to the original variable, and then runs another experiment. The student has an unconfounded set of experiments, though perhaps a less efficient one, but did not use VOTAT between every pair of trials. At the same time, we would not want to credit a student who ran hundreds of trials and through exhaustion managed to hit every possible set of parameters in the simulation. Machine learning can develop rules that handle these special cases that align to expert intuition.

We studied the possibility of identifying appropriate inquiry strategies in a way that goes beyond simple rules such as VOTAT and can recognize a broader range of appropriate strategies in the context of an online learning system named *Inq-ITS* (formerly called *Science ASSISTments*), an ITS for scientific inquiry (Gobert, 2015), shown in Figure 2. Within Inq-ITS, students make hypotheses, manipulate simulations and collect data, and then interpret the results of their experiments in terms of their original hypotheses. As shown in Figure 2, students typically manipulate simulations by changing the values of a small set of parameters, where the choices are categorical (i.e., 3 choices per variable) rather than continuous.

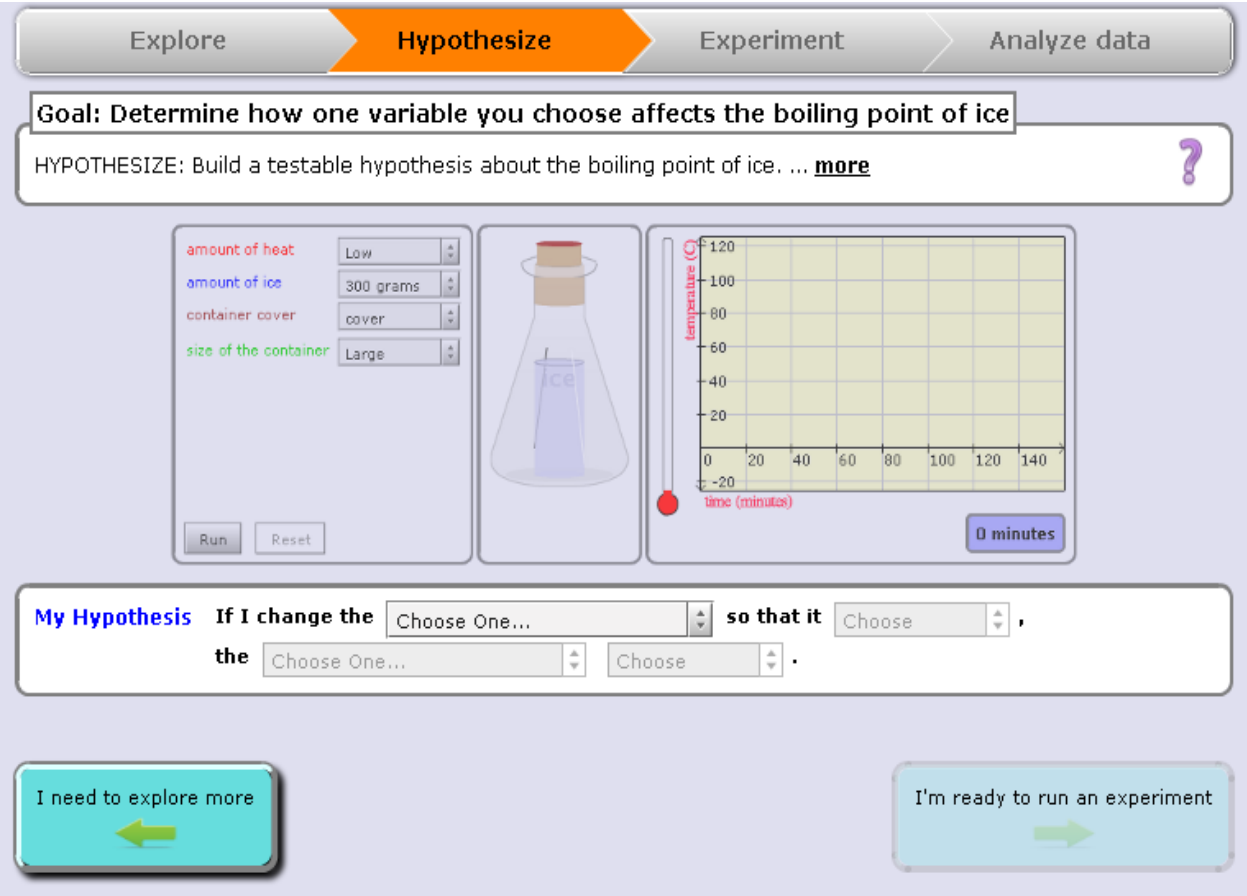


Figure 2. The version of Inq-ITS providing the data discussed in this article (more recent versions can be seen at http://www.inqits.com). Students make hypotheses, manipulate simulations and collect data, and then interpret the results of their experiments in terms of their original hypotheses.

To develop models of student inquiry within Inq-ITS, student behavior logged within Inq-ITS is transformed into a set of visualizations called "text replays". Originally proposed in Baker, Corbett & Wagner (2006), text replays are "pretty-printed" representations of student behavior over time, designed to be feasible for domain experts to read and use to identify behaviors or strategies of interest. For example, Figure 3, drawn from Gobert et al. (2012), shows text replays for Inq-ITS that were examined to identify whether the student is designing controlled experiments, whether the student is testing the stated hypothesis, and other aspects of student inquiry behavior. Within Inq-ITS, a single text replay corresponded to the actions a student made between creating hypotheses and interpreting their data in light of those hypotheses for a specific simulation. As Figure 3 shows, the human coders are able to see the relative time of student actions, what hypotheses the student made, what variables were manipulated between runs of the simulation to conduct experiments, and when (and how many times) the student ran the simulation. Not shown (but

also included in the text replays) are additional behaviors such as pausing the simulation, re-viewing the list of hypotheses, or opening the data table.

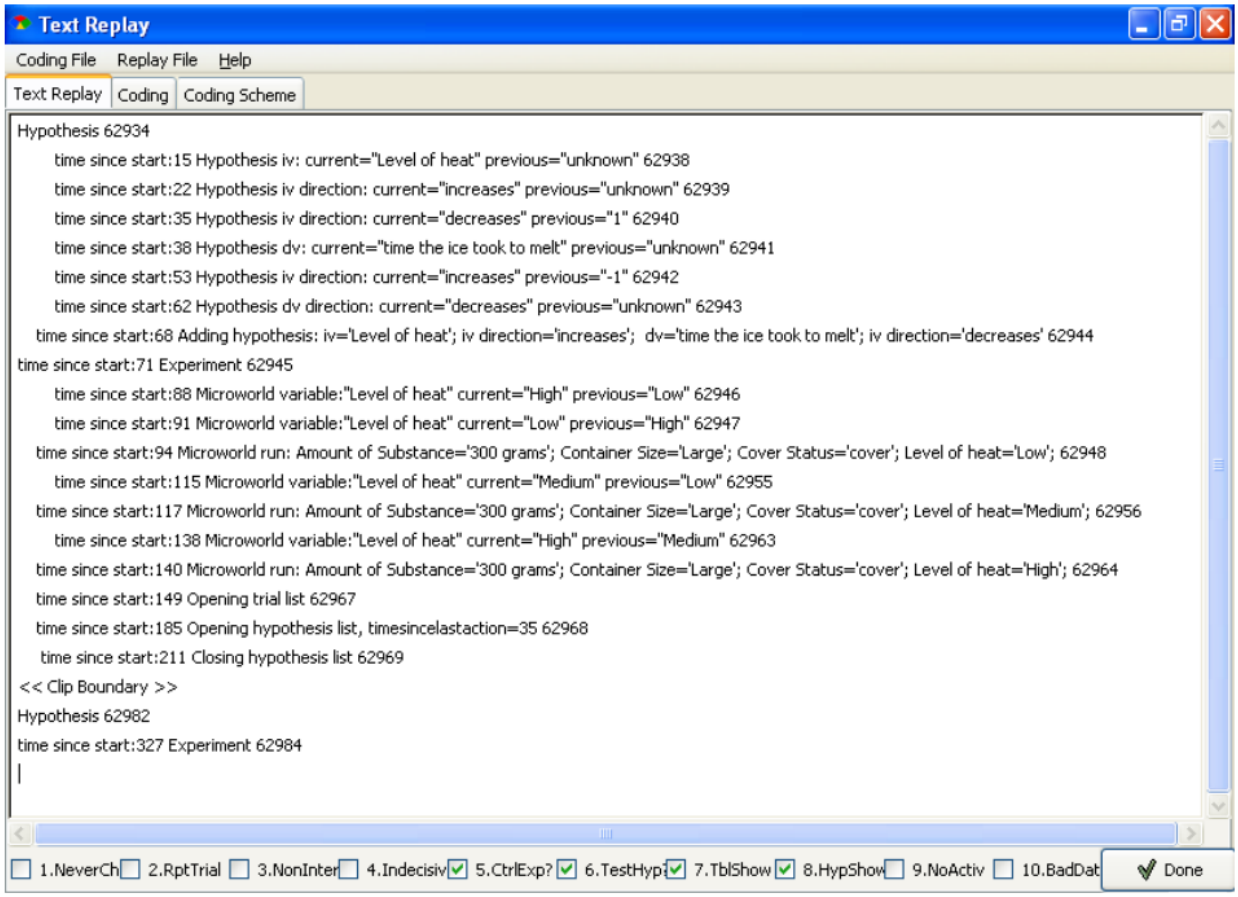


Figure 3. A text replay (readable depiction for data labeling) of student behavior in Inq-ITS.

The text replay infrastructure automatically samples which instances of student behavior will be coded by humans, in this case, stratifying the sample across students and across simulations (Sao Pedro et al., 2013a). Multiple coders label a sample of student behavior with reference to the constructs of interest and are checked for inter-rater reliability (San Pedro et al., 2013a). In this case, inter-rater reliability was good for labeling designing controlled experiments (Cohen's Kappa = 0.69) and perfect for testing the stated hypothesis (Cohen's Kappa = 1.00). A total of 571 sequences of student behavior were coded.

Next, a range of features of the student's interaction with the system was extracted from the raw data, including aspects of behavior such as how many times the student changed variables, the time between variable changes, repeated trials, and the degree of change between runs of the simulation. Features were filtered based on a domain expert's perception of which features would be most useful (Sao Pedro et al., 2012). The combined data set, including both features (as predictors) and labels made within text replays (as predicted variables), was then used as input to data-mining software. Multiple algorithms were tested to determine which made the best inference of the labels provided by the human coders from the predictor features. The algorithms were tested on data from entirely new students (Sao Pedro et al., 2013a), across the full range of contexts of use of the algorithm within the system (Sao Pedro et al., 2013a). The algorithms were also tested for validity within entirely different scientific domains. In this second validation, a model developed for a physical science simulation with simple relationships between the variables was validated

to work correctly when used in a biological science simulation, which contained more complex relationships (Sao Pedro et al., 2014). In the case of Inq-ITS, the algorithm that performed best was a relatively conservative decision tree algorithm, an unsurprising outcome given the relatively small data set. Overall, the model was able to distinguish appropriate science inquiry behavior from inappropriate science-inquiry behavior in a sequence of experimental trials 85% of the time.

These models functioned at the level of identifying whether a student demonstrated appropriate scientific-inquiry behavior in a single use of a simulation. They were also aggregated into broader inferences on student science-inquiry skill across simulations (Sao Pedro et al., 2013b, 2014). The broader models can then predict whether the student will be able to demonstrate the inquiry skill successfully in new situations. The model chosen for this was Bayesian knowledge tracing (BKT), a commonly used model for tracking student knowledge as it is changing (Corbett & Anderson, 1995). BKT takes a set of cases, each of which indicates whether a student is successfully demonstrating a skill, and aggregates the evidence over time into a running estimate of whether the student knows the skill and is likely to be able to demonstrate it in the future. The outputs from the inquiry behavior models were input into a BKT model, resulting in a model that could predict correctness on future simulations, as well as on future paper-based assessments of student inquiry skill (Sao Pedro et al., 2013b). This model was tested on new students and also across simulations (i.e., performance on one simulation was used to predict performance on another simulation – e.g., Sao Pedro et al., 2014). The model was able to predict future correctness for new students, in the same simulation, 74% of the time, and in different simulations 75% of the time.

Classification Based on Overall Inquiry Success

Our second example of inquiry modeling through machine learning involves a scenario in which it is not feasible to directly identify good and bad inquiry. Instead, we use external evidence that inquiry was successful as the basis for applying supervised machine learning. Specifically, in this example, we obtained data from students using a complex inquiry-learning environment where they had to answer a driving research question by collecting and interpreting data. In this case, our assumption is that the correctness of a student's answer to the research question would be highly correlated with productive inquiry behavior. Thus, we can use the scores from the students' submitted answer for labels, as the human-coded labels were used in the previous example, and analyze the data of student interaction with the environment to determine which student behaviors were associated with correct final answers.

This approach is used in work to model inquiry in virtual performance assessments (VPAs), which teach scientific-inquiry skills by presenting students with an authentic science problem in a virtual, simulated context (Clarke-Midura, Dede & Norton, 2011), shown in Figure 4. VPA provides students with the opportunity to interact with different scenarios in an immersive, three-dimensional (3D) environment. Students are asked to solve a scientific problem in each scenario by navigating an avatar through the virtual environment, making observations and gathering data. The avatar can collect several forms of evidence, conduct tests on it in a laboratory environment, talk to non-player characters, and read information kiosks.

Figure 4. VPAs, where students collect data, run experiments, talk to non-player characters, and read information resources in a rich, 3D virtual environment.

In this example, discussed in greater detail in Baker, Clarke-Midura, and Ocumpaugh (2016), we obtained data from a pair of VPA scenarios, referred to as the frog scenario and the bee scenario. In the frog scenario, students are asked to determine the cause of a frog mutation, specifically why the frog has six legs. The virtual world contains four farms with frogs and other evidence and where non-player characters provide competing opinions about the problem, a research kiosk where students can read about possible causes of the mutation, and a laboratory that contains dead, mutated frogs and lab water to use as comparisons. The student can collect tadpoles, frogs, and water samples from the farms, which they can then bring to the laboratory to run DNA and blood tests on the frogs and water-quality tests on the water samples. Possible explanations for why frogs are sick include parasites (the correct answer), pesticides, pollution, genetic mutation, and space aliens. Once students think they have collected enough information, they make a final claim for what they think is causing the frog mutation and support it with evidence. In the bee scenario, students are immersed in a similar environment, but in this context they are asked to determine what is killing off a local bee population (in this case, it is a genetic mutation).

Success in VPAs is identified by whether the student provides a correct final claim for why the frogs or bees are sick, and also by whether the student can correctly explain the chain of reasoning behind their claim. While it is possible to use rational or hybrid approaches to predict whether the student's behavior will be correct (e.g., Clarke-Midura & Yudelson, 2013); in this case, we engineered an automated model to predict whether the student behavior would be correct.

Our first step, as with Inq-ITS, was to distill a set of 48 semantically meaningful features (selected through structured brainstorming conducted by domain experts and machine-learning researchers) from the log files that contained data on all of the students' interactions with the environment. The features included behaviors such as moving from one virtual region to another, picking up or inspecting different objects, running laboratory tests on objects, reading informational pages at the research kiosks, and talking to non-player characters, as well as features regarding how long the student spent and details of the actions, such as how many different tests the students ran.

As in Inq-ITS, multiple algorithms were tested as potential mappings between the predictor features and classification labels. Algorithms were tested on data from entirely new students (Baker et al., 2016) to establish model validity. The models developed for each scenario were also validated to function effectively for the other scenario to increase confidence that these models would generalize to additional scenarios developed in the future for VPA. The algorithm that performed best was a relatively conservative decision rules algorithm (moderately more conservative than even the decision tree used in Inq-ITS). Overall, the model was able to distinguish appropriate science-inquiry behavior in a sequence of experimental trials 79% of the time. One surprising findings was that the time spent reading the various information pages was more predictive of the student eventually obtaining the correct answer than their specific exploratory or

experimentation behaviors in the system. Although surprising, this finding does not indicate that the virtual environment was not useful, but instead suggests that students may need declarative or expository information to best make sense of their activity within the virtual environment.

Pros and Cons

One of the biggest virtues of the machine-learning approach is that it can capture inquiry in complex settings where we don't know, a priori, exactly what good inquiry looks like. This is in contrast to the theoretically motivated approach discussed earlier. Many of the attempts to assess inquiry by more theoretical means require simplification of the inquiry task to allow inquiry skills to be recognized by clearly defined rules. While the theory-driven approach is well grounded, it risks incentivizing teachers and curriculum designers to create artificial tasks that teach students artificial specialized rules that cannot be used in many real-world inquiry tasks. To the extent that we want students to be able to use inquiry in murky, complicated, real-world situations, we need to create virtual environments where they learn inquiry strategies that are robust to situations and tasks that are not entirely straightforward.

Relatedly, this type of approach may be able to better distinguish inquiry strategies that are effective, even if not flawless, within simplified and rationally describable assessment tasks as well. Because the machine-learning approach is fundamentally data-driven, naturally occurring variations on successful strategies are more likely to be identified than in the case when content experts are required to define successful behaviors in advance. This allows machine learning to create models and assessments of student inquiry that apply more broadly and can be more flexible than purely theoretical approaches.

However, this flexibility carries with it limitations as well. Whereas a rule like VOTAT can be applied quickly across a broad range of learning systems, machine-learning models typically must be re-validated for different learning tasks, as was done above in both the Inq-ITS and VPA examples. In addition, machine-learned models are sufficiently complex that they can be harder to interpret by outsiders and do not always map back clearly to a theoretical understanding of the constructs. This can lead to questions of legitimacy and validity because it can be hard to prove that the models are not selecting for features that co-occur with successful inquiry within the given data set but are not actually relevant to good inquiry practice. Such possible spurious correlations could lead to models that validate within the original context but break down when used with a different population of students or systems. Additionally, machine learning can be more expensive than other approaches, both to create and validate models, and justify beyond the team that develops them.

Model-Driven Inquiry Detection

The previous two approaches focus on identifying specific feature markers in process data that indicate stronger or weaker application of inquiry or problem-solving skills. These methods stem from a scoring-oriented focus in which the motivating question might be: how would an expert rater recognize better applications of inquiry skills based on the data traces? In the model-driven approach, the focus shifts to the moment-by-moment decision making by the student. The motivating question here is quite different: how would we instruct an artificial agent to simulate the behavior of successful inquiry?

The model-driven approach attempts to create one or more generative models of within-task inquiry behavior based upon parameters that represent the latent traits we wish to make inferences about. The central feature of a model-driven approach is a mathematical model that does the following:

- links the latent-traits to be measured to the observable responses of the student,

- is based on a theory of how the latent-traits produce the responses, and

- predicts probabilities of specific responses as a function of both the person's latent traits and the response context (item or problem state).

Given such a model and sufficient performance data, the latent traits embedded in the model can be estimated using likelihood maximization methods. Given multiple competing models, each can be separately fit to the data and the inquiry performance can be classified based on the best fitting model. In this section, the approach is illustrated by a set of science-inquiry detectors formulated as Markov decision process (MDP) models that were developed to identify different inquiry strategies students applied to a chemistry simulation task.

Generative Models of Ideal Inquiry Behavior

The first step in the model-driven approach is to define models for the inquiry behaviors we wish to identify or evaluate. A single model of ideal behavior can be used as an expert model against which the student actions would be compared. In this case, we would estimate a single latent trait, which represents the student's ability to implement proper inquiry. Alternatively, multiple patterns of behavior can be modeled to provide more diagnostic information. Modeling both correct and incorrect strategies enables inference of not only how well the student is performing inquiry but also what misconceptions they might have or which strategies they might need to learn.

A probabilistic generative model will predict the probability of a student taking particular actions as a function of the current state of the simulation and the student latent traits. The model can be formulated to express the utility of each action as a function of the state of the experiment and the probability of choosing more useful actions as a function of the student ability. Because the action utility is dependent upon the current state of the simulation, the value of an action can change as students interact with the simulated experiment. For example, in a simulated experiment, we can imagine that a student has a choice of collecting more data under the current conditions, changing one of the experimental conditions, or analyzing the collected data. As the experiment progresses, collecting more data will become less useful and analyzing the data will become more useful. The utility of changing experimental conditions will depend upon the hypothesis being tested and the data that have already been collected.

One popular utility-based model is the MDP, which describes goal-driven behavior in a complex and possibly stochastic environment (Puterman, 1994). An MDP model relies on a definition of rewards, *R*, for achieving particular states along with costs for taking particular actions and a transition matrix *T*, which specifies the probability of transitioning from state *s* to *s'* given a particular action *a*. The reward structure *R* includes both a definition of the goal (the state which yields a high reward) and an encoding of motivation in the relative magnitudes of the goal reward and the cost of actions required to achieve the goal. The transition matrix *T* encodes beliefs about the problem space, in particular, giving the likely results of actions. MDP models are frequently used in the field of artificial intelligence for reinforcement learning (Barto, Sutton & Watkins, 1989) and have recently been used as psychometric models for estimating beliefs and ability from actions in complex tasks (Rafferty, LaMar & Griffiths, 2015; LaMar, under review).

MDPs for Inquiry Strategy Detection in the Concentration Simulation

This model-driven approach to inquiry assessment has been used with an experimental simulation-based assessment to infer inquiry skills based on student interactions with the simulation. The assessment involves mixing solvents and solutes and uses an embedded PhET simulation (Perkins et al., 2006), as shown in Figure 5. In a series of seven inquiry questions, students are asked about the relationship between amounts

of solute and resulting concentration in different mixtures. They are prompted to use the simulation to collect data and then are asked to respond and explain their response. Early cognitive labs showed that as middle school aged children interacted with the simulation they would run sequences of trials that corresponded to hypothesis-testing strategies. However, their strategies did not necessarily conform to the VOTAT-type strategy that the assessment designers had expected, nor did they confine themselves to a single strategy implementation. Instead a variety of strategies and strategy switching was commonly observed.

To identify the inquiry strategies used in particular interaction records while remaining resilient to occasional off-strategy behavior, MDP models were developed to embody both expected VOTAT strategies and additional strategies discovered in the initial data collection. Unlike the data-mining techniques described in the previous section, the generative-modeling approach requires all models to be defined in advance. The process of developing the candidate models, however, often include a combination of theoretical content knowledge and empirical discovery of student behaviors. For this study, information about student thought processes was gathered using cognitive lab protocols, in which students were asked to interact with the simulation and then explain their process to an interviewer. The combination of student-reported strategy use and the observed behavior of those students was analyzed in light of theory of student inquiry-skill acquisition to formulate preliminary strategy models. The models were then refined by using them to generate simulated student behavior and comparing the simulated actions to those taken by the original students and expert judgment of acceptable variation.

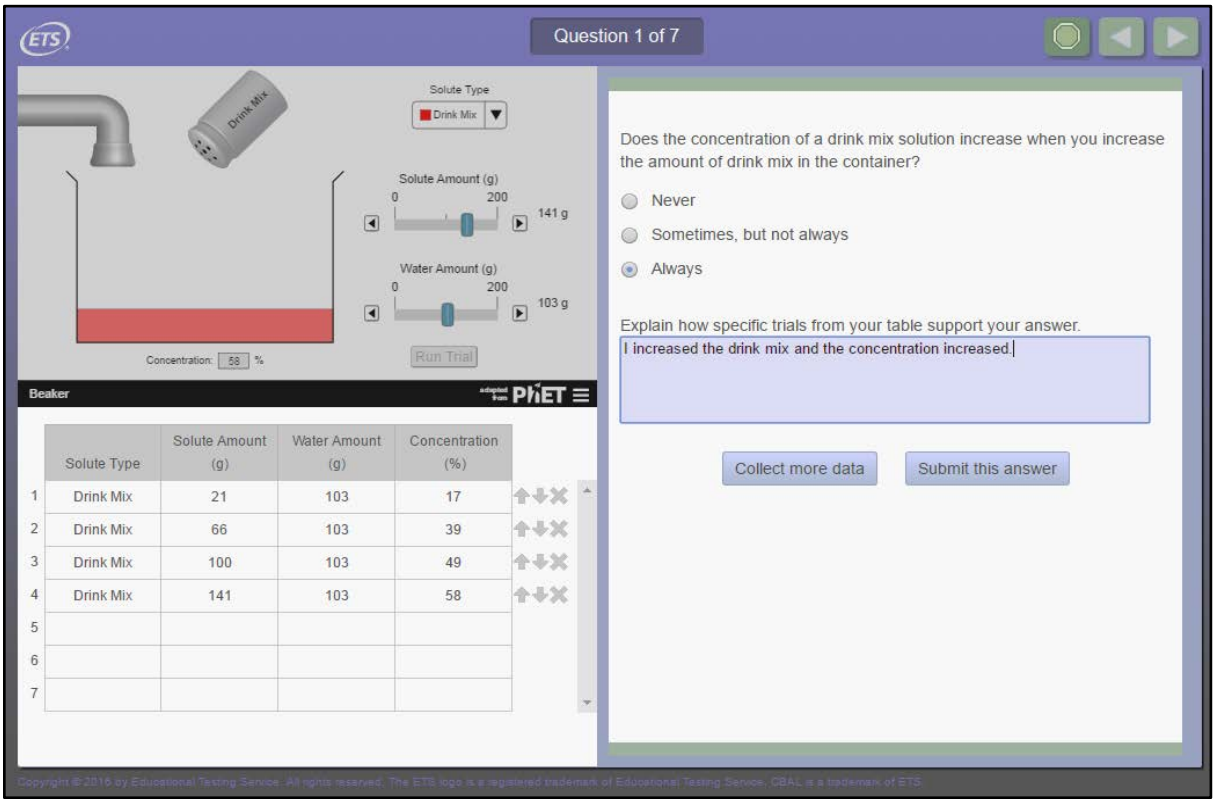


Figure 5. The second screen of item 1 for the concentration simulation.

To model inquiry behavior with an MDP, we needed to define the goal of the inquiry behavior, the sets of relevant actions and state variables, as well as the transition probabilities between states based on different actions. A single "ideal" inquiry model can be produced, which would allow comparison of student actions

with the expert model. This would result in a more score-oriented analysis where the student's inquiry skills can be estimated directly, assuming the expert model is the only valid strategy. Such an approach can be useful when the problem is fairly constrained such that there is a single correct goal with a preferred strategy for accomplishing the goal. The model would consider attempts to meet other goals as off-task behavior, resulting in a low estimate inquiry skill.

A more diagnostic approach can be taken by specifying different inquiry strategies as different MDP models. The models can then be fit to different sub-sequences of action data to estimate which strategies were being used at different parts of the inquiry process. Inquiry models can be developed for VOTAT strategies, directed search strategies, and other less productive approaches. As the goals and behaviors are quite specific, multiple productive and unproductive strategies can be modeled to provide both an overall inquiry score and diagnostic information that could be useful to adapt instruction.

Example Application of MDP Detectors

Based on data collected from 150 adult participants in an Amazon Mechanical Turk pilot study, three different inquiry strategies were identified and coded as MDPs for student interactions with question 1 (shown in Figure 5) and question 7, which was identical. Two of the strategies are different instantiations of the VOTAT strategy, one in which water is held constant and the solute is gradually increased (Increase Solute Strategy [ISS]) and the other in which water is held constant and the solute is gradually decreased (Decrease Solute Strategy [DSS]). For both ISS and DSS, the goal is to gather enough data to be able to answer the question. Because implementers of these strategies understand the importance of control-of-variables, data are considered to be a trial in which the amount of water is unchanged from the previous trial, but the solute amount is greater.

The third strategy involves a directed search in which students seek a saturated solution (Find Saturation Strategy [FSS]). The goal of students implementing FSS is to test whether the solution will indeed saturate. Students who implement this strategy are assumed to have a fair amount of content knowledge because they know that saturation is possible and they further know what conditions are likely to cause saturation. The typical behavior for this strategy involves setting the solute to a high level and dramatically decreasing the amount of water added until saturation is detected, often multiple times.

Note that all three of these strategies are productive inquiry strategies. While some non-productive behavior was observed based on student misconceptions, insufficient examples of those behaviors have been collected to formulate and test an appropriate MDP model.

To implement the identified inquiry strategies as MDPs, each strategy's goals were translated into a reward structure. Similarly, their beliefs were translated into action sets, state space variables, and transition functions. Table 1 shows the goals, rewards, and beliefs for ISS and FSS. The DSS strategy was similar to ISS with only data counted that decrease the solute rather than increase it.

Once the MDP generative models were built for the three different strategies, the models were fit to the record of trials run in the simulation to identify which strategies were most likely being used. For each student record for a particular item, the sequence of trials ran may contain zero, one, or more instances of a strategy implementation. To enable detection of strategy implementations at any point in the sequence and of any length, the trial sequences are split into all possible sub-sequences above the minimum length of 3 trials. Each of the inquiry strategy models is then fit to each candidate sequence, giving a likelihood that that particular sequence was generated by a student attempting to implement the inquiry strategy. Final strategy sequence labels were determined by maximizing likelihood over the entire record.

Table 1. Model components for the ISS and FSS models.

	Increase (Decrease) Solute Strategy	Find Saturation Strategy
Goal	Gather enough data to determine how increasing solute affects concentration at one water level.	To determine if this solution will saturate.
Rewards	Increase with: • More data • Range covered	Increase with: • Finding saturation • More data
Beliefs	Data = Two trials with an increase (decrease) in solute but constant water Water amount unimportant as long as it doesn't change	Data = Trial with high solute and low water Saturation = More than one concentration result of the same value

Using this method, log records from the first and last questions (1 and 7) were analyzed and sequences of implemented inquiry strategies were identified. For example, some students implemented multiple ISS strategies in a row (altering the amount of water in between), giving detected patterns of ISS-ISS-ISS, while others showed strategy switching such as ISS-FSS. Overall, the students who used more than one strategy implementation scored better on the following content question, indicating that more sophisticated inquiry patterns correlate with more successful conclusions.

Pros and Cons

There are a number of advantages of the model-based approach. As a theory-driven approach, the results are easily interpretable and can fit into existing frameworks for assessing science practice. Implemented as discrete single-strategy detectors, as demonstrated by the MDP detectors, multiple different inquiry strategies can be identified, including both productive strategies and those based on misconceptions, making the approach ideal for formative assessment and tutoring scenarios. The probabilistic nature of the models, meanwhile, allow for detection of strategy behavior even when the implementation of the strategy is imperfect. These factors make this approach useful in complex, open-ended inquiry tasks.

On the other hand, such complexity comes at a cost. The theory-driven modeling requires that theory exists to explain very low-level actions. Such fine-grained theory is often lacking in the current science literature. The formulation of MDP models also requires an understanding of how utility-theoretic state-space models work, making this method potentially difficult to access for science educators. Furthermore, the models that are developed are context-specific. While models for generalizable strategies, such as VOTAT, might be structurally similar in different tasks, each model needs to be customized to the actions and variables available in specific tasks.

Implications for GIFT and Tutoring

ITSs frequently include the affordance for simulation-based inquiry and problem solving (Murray, 2003). Principled assessment of student skills based on their actions within such tasks are critical to produce the relevant feedback and guidance that one expects from an advanced ITS.

For any of these approaches to be useful in a computerized tutoring environment, the identification and assessment of inquiry skills must be available in the moment, not merely in post-processing. MDP and BKT are not currently available in the GIFT framework, but their inclusion would be relatively straightforward.

In particular, future work to create flexible, lightweight versions of MDP algorithms could make them more accessible to tutoring systems for online processing. There is already support for the inclusion of automated detectors based on Rapid Miner into GIFT trainee models, making this type of algorithm readily usable within the GIFT framework. Theoretically defined behavioral indicators, meanwhile, can be easily coded into the logic of the task programming because their definition goes hand in hand with the task design.

Conclusions

Computers allow students to interact with more complex science and problem-solving scenarios and educators call for teaching and assessment to include more realistic science and engineering tasks. As a consequence, identifying and assessing inquiry strategies has become both pertinent and possible. This chapter has outlined three different approaches to assessing the science practice of hypothesis testing: a theory-driven approach based on carefully crafted assessment tasks and corresponding performance indicators, a more theoretical approach based on using machine learning to identify successful or appropriate inquiry behavior, and an approach that builds off theoretical understanding but uses generative process models to recognize a wider range of behavior.

Each of the approaches have their advantages and limitations, with the most appropriate approach likely dependent upon the type of task presented to the students and the types of skills intended to be taught and assessed. The theory-driven behavioral indicators are likely the best choice for high-stakes assessment because the behaviors detected are well understood and scoring is clean and defensible. The machine-learning approach provides a method for assessing inquiry in wide open environments and situations in which good inquiry cannot be cleanly defined. The model-based approach is something of a compromise between the pure theoretically defined behavioral indicators, which rely on consistently predictable behavior patterns, and the machine-learning inquiry detectors, which can discover patterns of behavior that experts might not predict. The MDP inquiry detectors are based in theory, although the theory can be developed iteratively with qualitative analyses of existing log files. Their probabilistically framed model allows strategy behaviors to be detected even when the strategy is imperfectly implemented.

References

Baker, R., Clarke-Midura, J., Ocumpaugh, J. (2016) Toward General Models of Effective Science Inquiry in Virtual Performance Assessments. Journal of Computer Assisted Learning, 32 (3), 267–280.

Baker, R.S.J.d., Corbett, A.T., Wagner, A.Z. (2006) Human Classification of Low-Fidelity Replays of Student Actions. *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems*, 29–36.

Baker, R., Siemens, G. (2014) Educational data mining and learning analytics. In Sawyer, K. (Ed.) *Cambridge Handbook of the Learning Sciences: 2nd Edition*, pp. 253–274.

Barto, A. G., Sutton, R. S. & Watkins, C. J. C. H. (1989). Learning and Sequential Decision Making. In *LEARNING AND COMPUTATIONAL NEUROSCIENCE* (pp. 539–602). MIT Press.

Clarke-Midura, J., Dede, C. & Norton, J. (2011). Next generation assessments for measuring complex learning in science. *The Road Ahead for State Assessments*, 27–40.

Clarke-Midura, J. & Yudelson, M. (2013) Towards Identifying Students' Reasoning using Machine Learning. *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, 704–707.

Corbett, A.T., Anderson, J.R. (1995) Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction, 4, 253-278.Dede, C. (2010). Comparing frameworks for 21st century skills. *21st Century Skills: Rethinking How Students Learn*, *20*, 51–76.

Dede, C. (2010). Comparing frameworks for 21st century skills. *21st Century Skills: Rethinking How Students Learn*, *20*, 51–76.

Gobert, J. D. (2015). Inq-ITS: design decisions used for an inquiry intelligent system that both assesses and scaffolds students as they learn. *Handbook of cognition and assessment. New York: Wiley/Blackwell.*

Gobert, J.D., Sao Pedro, M.A., Baker, R.S.J.d., Toto, E., Montalvo, O. (2012) Leveraging Educational Data Mining for Real-time Performance Assessment of Scientific Inquiry Skills within Microworlds. Journal of Educational Data Mining, 4 (1), 111–143.

Greiff, S., Wüstenberg, S. & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92–105.

Greiff, S., Wüstenberg, S. & Funke, J. (2012). Dynamic problem solving: a new measurement perspective. *Applied Psychological Measurement, 36*, 189–213.

Jonassen, D. H. (Ed.), *Learning to solve complex scientific problems*. New York: Lawrence Erlbaum.

Klahr, D. (2002). *Exploring science: The cognition and development of discovery processes.* Cambridge: MIT press.

Klahr, D. & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1–48.

Kuhn, D. (2012). The development of causal reasoning. *WIREs Cognitive Science, 3*, 327–335.

Kuhn, D., Black, J., Keselman, A. & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction, 18*, 495–523.

LaMar, M. (under review). Markov decision process measurement model. Manuscript under review.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L. & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233.

Murray, T. (2003). An Overview of Intelligent Tutoring System Authoring Tools: Updated analysis of the state of the art. In *Authoring tools for advanced technology learning environments* (pp. 491–544). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-94-017-0819-7_17.

NGSS Lead States. (2013). The Next Generation Science Standards: For States, By States. Achieve, Inc. on behalf of the twenty-six states and partners that collaborated on the NGSS. Retrieved from http://www.nextgenscience.org/next-generation-science-standards.

OECD (2014). *PISA 2012 results. Creative Problem Solving*. Paris: OECD.

Paquette, L. & Baker, R.S. (under review) Comparing Machine Learning to Knowledge Engineering for Modeling SRL Behaviors: A Case Study in Gaming the System. Manuscript under review.

Perkins, K., Adams, W., Dubson, M., Finkelstein, N., Reid, S., Wieman, C. & LeMaster, R. (2006). PhET: Interactive simulations for teaching and learning physics. *The Physics Teacher*, *44*(1), 18–23.

Popper, K. R. (1959). The logic of scientific discovery. *London: Hutchinson*.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. New York: John Wiley & Sons, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=528623.

Rafferty, A. N., LaMar, M. M. & Griffiths, T. L. (2015). Inferring Learners' Knowledge From Their Actions. *Cognitive Science*, *39*(3), 584–618. https://doi.org/10.1111/cogs.12157.

Sao Pedro, M., Baker, R.S.J.d., Gobert, J. (2012) Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012)*, 249–260.

Sao Pedro, M. A., Baker, R. S. & Gobert, J. D. (2013a). What different kinds of stratification can reveal about the generalizability of data-mined skill assessment models. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 190–194). ACM.

Sao Pedro, M.A., Baker, R.S.J.d., Gobert, J., Montalvo, O. Nakama, A. (2013b) Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction, 23* (1), 1–39.

Sao Pedro, M., Jiang, Y., Paquette, L., Baker, R.S., Gobert, J. (2014) Identifying Transfer of Inquiry Skills across Physical Science Simulations using Educational Data Mining. *Proceedings of the 11th International Conference of the Learning Sciences*.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76. https://doi.org/10.1016/j.asw.2013.04.001.

Vollmeyer, R., Burns, B. D. & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75–100.

Wieman, C. (2015). Comparative cognitive task analyses of experimental science and instructional laboratory courses. *The Physics Teacher*, *53*(6), 349–351.

Wüstenberg, S., Greiff, S. & Funke, J. (2012). Complex problem solving. More than reasoning? *Intelligence, 40*, 1–14.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172–223.

CHAPTER 13 – Automated Assessment of Learner-Generated Natural Language Responses

Vasile Rus[1], Andrew M. Olney[1], Peter W. Foltz[2], and Xiangen Hu[1]
The University of Memphis[1], Pearson, Inc. and University of Colorado[2]

Introduction

Eliciting open-ended learner responses gives learners the opportunity to freely generate a response to specific prompts as opposed to discriminating among several predefined responses (one of which is correct). In the latter case, i.e., multiple-choice questions, there is always the risk of learners picking the correct answer for the wrong reasons without the possibility of identifying such flawed knowledge. Open-ended responses eliminate such risks, reveal students' thought processes, document explanation of reasoning, and allow for the expression of creative and original correct responses, none of which is possible when using multiple-choice assessment instruments (Parmenter, 2009; Walstad & Becker, 1994). Therefore, freely generated student responses to a question, e.g., asked by an intelligent tutoring system (ITSs; Rus et al., 2013), or an essay prompt, e.g., as in the high-stake exam Scholastic Assessment Test (SAT), should be used in educational contexts because they provide a unique opportunity to assess students' knowledge and skills in an area.

The Need for and Challenges with Automated Assessment of Learner-Generated Natural Language Responses

The challenges (cost and effort) that arise from manually assessing open-ended student responses limit their use by educators. Automated methods to assess student-generated responses are fervidly pursued to address the cost and effort challenges. It should be noted that automated methods have the additional advantage of systematically and consistently assessing student responses as compared to human raters. This chapter offers an overview of methods for automatically assessing students' freely generated answers.

The self-generation process, the key feature of open-ended assessment, offers unique opportunities and challenges for automated assessment. An effect of the self-generation aspect of open-ended responses, which is an advantage and a challenge at the same time, is their diversity along many quantitative and qualitative dimensions. For instance, free responses can vary in size from one word to a paragraph to a document. The challenging part is the fact that there needs to be a solution that can handle the entire variety of student responses, a tall order. Indeed, analyzing students' responses requires natural language techniques that can accommodate the endless variety of student input and assess such input accurately.

Another major challenge is that open-ended responses may be assessed in different ways that depend on the target domain and instructional goals. This makes it difficult to compare assessments. For example, in automated essay scoring the emphasis is more on how learners argue for their position with respect to an argument the essay prompt while in other tasks, such as conceptual physics problem solving, the emphasis is more on the content and accuracy of the solution articulated by the learner. This chapter provides an overview of the opportunities, challenges, and state-of-the-art solutions in the area of automated assessment of learner-generated natural language responses.

Automated Methods For Assessing Learner-Generated Natural Language Responses

We focus in this chapter on natural language open-student responses as opposed to responses that require, for instance, drawing a diagram. Assessing such natural language learner responses is a natural language understanding (NLU) task.

Computational approaches to NLU can be classified into three major categories: true-understanding, information extraction, and text-to-text similarity. In true understanding, the goal is to map language statements onto a deep semantic representation such as first-order logic (Moldovan & Rus, 2001; Rus 2002; Banarescu et al., 2013; Bender, Flickinger, Oepen, Packard & Copestake, 2015; Bos, 2015; Liang, Jordan & Klein, 2013). This approach relates language constructs to world and domain knowledge that is stored in a well-specified computational knowledge base (Lenat, 1995), and that ultimately enables inferencing. Inconsistency and contradictions can be automatically detected, revealing potential flaws in students' mental model of the target domain. Current state-of-the-art approaches that fall into this true-understanding category offer adequate solutions only in very limited contexts, i.e., toy-domains. These lack scalability and thus have limited use in real-world applications such as summarization or ITSs. Notable efforts in ITSs are by VanLehn and colleagues (Rosé, Gaydos, Hall, Roque & VanLehn, 2003; VanLehn, Jordan, Rosé, et al., 2002). Related efforts use less expressive logics like description logic (Zinn, Moore, Core, Varges, Porayska-Pomsta, 2003) in conjunction with the two other NLU approaches' categories discussed next.

Information extraction approaches use shallow processing to automatically detect in learners' free responses the presence of certain nuggets of information that represent key expected concepts or derived measures that could be used as predictors of student responses' correctness or overall quality. These approaches focus on text surface features or matching of exact words. They are most appropriate for item types such as fill-in-the-blank short answers where there is a limited range of expected correct responses.

Text-to-text (T2T) similarity approaches to textual semantic analysis avoid the hard task of true understanding by defining the meaning of a text based on its similarity to other texts, whose meaning is assumed to be known. Such methods are called benchmarking methods because they rely on a benchmark text, which is generated, checked, or annotated by experts, to identify the meaning of new, unseen texts. We focus in this chapter primarily on T2T approaches because they are the scalable and dominant practical approaches at this moment and probably for the foreseeable future. One may argue that some of the information extraction approaches fall under the category of T2T approaches because, for instance, identifying in student responses a number of expected concepts that are specified by an expert is equivalent with a T2T approach in which parts of student responses are compared to key expected concepts provided by experts. To make the distinction less ambiguous for such approaches, we call T2T approaches those that compare a full student response to a full benchmark or ideal response provided by an expert. If the expert only provides a set of key expected concepts, then such an approach falls under the information extraction category.

We also distinguish between two major categories of student responses: argumentative essays versus accurate-content responses, i.e., short essays. While these two types of essays are not mutually exclusive, nor are they inclusive of all essay types, these types of student responses are the result of two different major instructional goals. In argumentative essay scoring, there is an emphasis on *how* learners argue for their position with respect to the essay prompt, whereas in a task that scores conceptual physics problem solving the emphasis is on *what* is said, i.e., content/accuracy of the solution articulated by the learner. It should be noted that content is also considered in automated scoring of argumentative essays as, for instance, the essay must be "on topic", i.e., related to the essay prompt. Nevertheless, in this case factors such as vocabulary, grammaticality, and argumentation (the *how*) are emphasized. On other other hand, grammaticality and vocabulary size/richness are less important aspects to consider while assessing the correctness of physics problem solutions as compared to the conceptual correctness of the articulated solution. While grammaticality is not critical when assessing for correctness, it could be important because often content

knowledge cannot be fully explicated without correct expression. For instance, a response with a bad grammar will negatively impact automated syntactic parsing, which in turn will negatively impact the overall outcome of the automated assessment method. Bad grammar leads to bad parsing and so on, and it leads to error propagation throughout the automated assessment software pipeline. However, from our experience, we can testify that syntactic information does not add significant more value on top of other factors for predicting accuracy.

This chapter provides an overview of T2T opportunities, challenges, and state-of-the-art solutions. It is by no means a comprehensive review of previously published work in this area. Rather, it is a short summary of the area.

It is important to add one meta-comment regarding terminology before proceeding further with details of various T2T methods for automated assessment of learner-generated natural language responses. The authors of this contribution do not adhere to the so called "constructed student answers" label, which has been used by some researchers when referring to learner-generated natural language responses (Leacock & Chodorow, 2003; Magliano & Graesser, 2012). First, this label implies that students construct their answers from some predefined primitives/blocks, e.g., pieces of the response that might be available to them. This is not the case as students freely generate the answers. Second, the "constructed" terminology has its roots on the fact that experts construct a reference/model/ideal answer, which is then used to assess student responses. That is, the "constructed student answer" terminology links the nature of students' answers to the evaluation method (Leacock & Chodorow, 2003; Magliano & Graesser, 2012), i.e., comparing the student answer to a constructed reference/model/ideal answer provided by experts. We believe that the way students generate the answer, e.g., freely composing the answer versus selecting it from a set of forced answer choices versus something else, should be the only aspect that should inform the choice of a label of what is being assessed. For instance, a freely composed student answer can be assessed by a true understanding method as opposed to a semantic similarity approach that relies on a reference answer or reference set of concepts.

Automated Essay Scoring

Automated essay scoring (AES) has become increasingly accepted with multiple systems available for implementing the scoring of writing for learning applications and ITSs (Shermis & Burstein, 2013). Studies of AES systems have shown that they can be as accurate as human scorers (e.g., Burstein, Chodorow & Leacock, 2004; Foltz, Laham & Landauer, 1999; Landauer, Laham & Foltz, 2001; Shermis & Hamner, 2012), can score on multiple traits of writing (Foltz et al., 2013), can provide accurate feedback on content (Beigman-Klebanov et al., 2014; Foltz, Gilliam & Kendall, 2000), and can score short responses (Higgins et al., 2014; Thurlow, et al., 2010). These systems are now also being used operationally in a number of high stakes assessments like General Educational Development (GED) and various state K–12 assessments (Williamson et al, 2010), placement tests like Pearson Test of English and ACCUPLACER (suite of tests that assess reading, writing, math, and computer skills), and for writing practice in systems like Criterion and WriteToLearn.

While there are differences among the various AES systems and the methods they employ, most share a common approach. This approach can be simply described as 1) represent a student's writing quantitatively as a set of features and 2) determine how to weigh and combine the features to best characterize the quality of the student writing. The features are quantifiable natural language processing features that measure aspects such as the student's expression and organization of words and sentences, the student's knowledge of the content of the domain, the quality of the student's reasoning, and the student's skills in language use, grammar, and the mechanics of writing. In creating these features, it is critical that the computational measures extract aspects of student performance that are relevant to the constructs for the competencies of

interest (e.g., Hearst, 2000; Williamson, Xi & Breyer, 2012). For example, and explicated in greater detail in the following, features that measure the semantic content of student writing are used as measures of the quality of a student's domain knowledge. A measure of the type and quality of words used by a student provides a valid measure of lexical sophistication. However, measures that simply count the number of words in an essay, although it may be highly correlated with human scores for essays, do not provide a valid measure of writing sophistication.

To score a student-generated essay, multiple language features are typically measured and combined to provide a score or feedback. Combining and weighing the derived features is then performed by learning to associate the features with examples of student performance. These examples of performance can include samples of student responses that have been pre-scored by human raters or examples of ideal responses created or selected by experts. Machine learning (e.g., regression, Bayesian, classification) techniques are then used to build a model that weighs the features in relationship to the examples of performance while maximizing predictive performance and generalizability. The details of these approaches are beyond the scope of this chapter, but are covered in detail in Shermis and Burstein, (2013). However, we do describe important considerations in the features and methods used for assessing student-generated essays, most particularly with respect to content accuracy.

The recent Automated Student Assessment Prize (ASAP), funded by the Hewlett Foundation, revealed that current AES technologies rival human grader's performance (Shermis & Hamner, 2012; Shermis, 2014), although there are debatable aspects of the used methodology (Perelman, 2013). The ASAP exercise, which drew participation from seven commercial organizations and one academic lab, has shown that the dominant approach is a form of semantic similarity in which new, to-be-scored essays are compared to human-graded essays whose quality is already known.

Accurate-Content Essay Scoring

When assessing learners' responses for content accuracy or correctness, the most widely used approach is semantic similarity. As already mentioned, such an approach involves experts providing one or more ideal responses to specific prompts or sets of student responses that are pre-scored by experts. New student responses to the same prompts are subsequently compared to these ideal or benchmark responses. The main advantage of the semantic similarity approach is the circumvention of the need to acquire and automatically use world and domain knowledge explicitly, which are required for a true understanding of learners' natural language responses. It is assumed and hoped that the ideal response provided by experts contains enough cues that allows a simple or augmented semantic similarity process to make a good enough judgment on the correctness of students' responses.

While a semantic similarity approach still requires experts to generate ideal/benchmark responses, or score previously collected student responses, this manual step leads, overall, to a more scalable and cost-effective approach than the true understanding approaches. Consequently, the semantic similarity approach currently dominates.

Before presenting more details about such semantic similarity approaches to assessing the accuracy/correctness of learner responses, we present key issues with automatically assessing students' answers using a semantic similarity approach. Given the focus of this volume, we focus on student answers that need to be assessed in the context of intelligent interactive systems, namely, ITSs. In these systems, learners are prompted to provide answers and these answers must be assessed in real time to provide appropriate feedback and trigger appropriate instructional strategies that maximize learning.

Key Challenges

There are many challenges when it comes to automatically assessment of students' open-ended responses. We start by enumerating some of these challenges.

Spelling and Grammatical Errors. Students' responses often contain spelling and grammatical errors. Depending on various learners' verbal and writing abilities, they will generate responses that vary in compositional quality. The good news is that reasonably good spelling and grammar correction software tools are available, at least for English. Furthermore, since the focus on content-accuracy assessment is often on content, issues such as spelling and grammatical errors are not a major conceptual hurdle to overcome. That being said, there are pragmatic implications due to the current state of natural language technologies, e.g., modest syntactic parsing accuracy, as mentioned earlier.

Size Variability. There is great variability in student responses when it comes to size. For instance, student responses can be a single word, a phrase or chunk, a segment of a sentence such as a clause or several clauses, a complete sentence, a paragraph/short-essay or even a full essay. One needs to develop an approach that can handle student responses of various granularity or, alternatively, develop a separate approach for each of the mentioned granularity levels. Some approaches can be easily extended from a word to a sentence level but not to multiple-sentences/paragraph level. Other approaches scale well from words to sentences to paragraphs at the expense of disregarding information that is important for precise assessments (Rus, Banjade & Lintean, 2014).

Dialogue Utterances. Specific challenges arise when assessing learners' language in a multi-turn dialogue context such as in dialogue-based ITSs. For example, elliptical responses similar to response #5 in Table 1 are quite frequent. In spoken dialogue, the challenges are even greater due to peculiarities of spoken dialogue, which we address later in the spoken versus written input section.

Heavily Contextualized. In many situations students' responses are highly contextualized. For example, all the student answers shown in Table 1 are in the context of one particular physics problem that was used in an experiment with high school students who learned physics from a state-of-the-art ITS called DeepTutor (Rus, D'Mello, Hu. & Graesser, 2013). All these student responses must therefore be assessed while accounting for the physics problem, which is the focus of the dialogue. Furthermore, the responses needed to be interpreted in the context of the full dialogue history up to the point where the response was generated. In general, the responses in Table 1 are targeted responses to a previous tutor question, which is yet another contextual dimension.

Table 1. Examples of actual high school student answers showing the diversity of student responses to various prompts from the state-of-the-art intelligent tutoring system DeepTutor.

ID	Student Responses
1	The force exerted by gravity and tension of the rope are equal.
2	These forces balance each other.
3	The tension is equal to the force of gravity.
4	They are equal.
5	Equal.
7	The tension in the rope is greater than the downward force of gravity.
8	The tension in the rope is greater than gravity in order to raise the child upward.
9	They are equal and opposite in direction.
10	The tension in the rope is equal to the mass of boy times gravity. Newton's second law states the force is equal to mass times acceleration. In this case, the tension is the force. Gravity is the acceleration.

Textual responses vs. multi-modal. This chapter focuses on textual student responses but multimodal responses that contain text, diagrams, and other non-textual products, e.g., a diagram, are often generated and need to be assessed. There are specific challenges in these cases such as aligning the textual and non-textual elements to generate a more complete model of the student response. Even when student responses are pure textual, there may be reference to non-textual elements provided in context, e.g., a physics problem usually has a picture attached to it describing a visual rendering of the problem scenario. Students may refer to the provided picture instead of simply focusing on the concepts mentioned in the textual description of the problem. For instance, students may simply say "*The truck is pushing the car to the right*" when no explicit spatial relationship was mentioned in the problem description or the previous dialogue with the ITS; such spatial information, although irrelevant to answering the main question of the problem, is conveyed through the accompanying image depicting a truck pushing a car.

Core Linguistics Issues. Additional key linguistic issues often need to be addressed for a more comprehensive solution even when a semantic similarity approach is being used. These key linguistics tasks include coreference and anaphora resolution as students often refer to entities mentioned earlier in the dialogue and problem description using, for instance, pronouns (Niraula, Rus, Stefanescu, 2013; Niraula, Rus, Banjade, Stefanescu, Baggett, and Morgan, 2014), negation (Banjade, Niraula, and Rus, 2016), and synonymy (different words with same meaning) and polysemy (many meanings of the same word). Some of the semantic similarity approaches we summarize do include simple and sophisticated solutions for word sense disambiguation to address, for instance, polysemy. Semantic methods such as Latent Semantic Analysis (LSA; Landauer, Foltz & Laham, 1998) handle synonymy indirectly by mapping words into a semantic space where synonym words will be close to each other.

Written vs. Spoken Input. Students' responses can be typed or spoken. Spoken language has two main challenges: register versus technical. There are specific phenomena that are more prevalent in spoken language than written language. For example, there are disfluencies such as stop-restart segments where a speaker starts an utterance and suddenly stops before finishing it to restart and utter a replacement. Concrete examples of such disfluencies are fillers ("uh", "um"), which are dialogue specific particles with no particular content value, false starts/re-starts ("I I really wanted …"), repetitions ("I would like a coffee a hot coffee"), and corrections ("I would like a an apple").

Compositional Approaches

We focus in this section on compositional approaches to addressing the task of automated assessment of student open responses. The *principle of compositionality* states that the meaning of a text can be determined by the meaning of its constituents and the rules used to combine them (Lintean & Rus, 2012). The major steps in compositional approaches are outlined as follows:

- *Step 1. Derive a word-level meaning representation.* This might include some necessary preprocessing of the input text, which could include the detection of other constructs such as a noun phrase or a verb phrase.

- *Step 2. Discourse level processing.* This includes resolution of content references, such as pronouns to their referents, to maximize the outcome of the next step, i.e., the alignment step.

- *Step 3. Alignment.* Align words (or other constituents) across the paired student vs. ideal responses and combine the word-level meaning representations. The alignment could be based on lexical information, i.e., words and their meanings, and also take into account relations among words such as syntactic relations. Other contextual elements such as the nearby words (within a window of, say, three words before or three words after) or directly related words via a syntactic dependency could be considered. The alignment could also include explicit negation particles.

- *Step 4. Report an overall, usually normalized, similarity score.* The resulting score could be sanctioned by extra semantic elements such as negation focus and scope. For instance, Rus and Graesser (2006) altered the alignment score using a negation term that accounted for single or double negation.

- *Step 5. Map the similarity score into qualitative decisions*. After obtaining the similarity score a qualitative decision is being derived such as the student response is correct or the student response is incorrect.

Word-level and Text-level Meaning Representations

We focus in this section on the first step of the semantic similarity procedure outlined previously and which consists of deriving meaning representations of texts, a key issue in NLU. As previously stated, T2T approaches define the meaning of a text based on its similarity to other texts.

Perhaps the simplest approach is to represent texts as sets of words and as a similarity metric the set intersection of the words of the two texts under consideration, i.e., $U \cap V$ for texts U and V. This intersection can be normalized by the cardinality or union of U and V to yield a variety of set-based metrics initially proposed by Paul Jaccard and known as the Jaccard similarity coefficient. Set-based metrics have the advantage that the calculation of similarity between texts is fairly direct (Manning & Schütze, 1999). In this family of "set-of-words" approaches, words are not given any particular kind of representation but texts are treated as sets of words. As we explain later, even for such methods of set-based textual semantic representations and similarity approaches, there is an underlying 1-of-N vector representation with binary weights (as opposed to representations based on distributional properties of words derived from large collections of natural language texts, which are explained next).

A more sophisticated family of approaches is based on the distributional vector space representation in which the meaning of a word is represented as a vector in a multi-dimensional space. The dimensionality of the space varies from one vector space representation to another. Furthermore, these distributional vector space approaches rely on the distributional hypothesis according to which the meaning of a word is defined by the company it keeps. That is, they derive the vector representation of words starting with a word-to-word co-occurrence analysis over a large collection of texts, as detailed later. Originally conceived for information retrieval, where the purpose is to match a textual query with the most relevant or similar document in a large collection of documents (Salton, Wong & Yang, 1975), distributional vector space approaches are perhaps the most widely used for T2T similarity.

As already mentioned, distributional vector representations rely on a statistical analysis of word co-occurrences in large collections of natural language texts, i.e., a corpus such as Wikipedia articles (Ştefănescu, Banjade & Rus, 2014). A typical and simple co-occurrence analysis consists of generating a term-by-document matrix where each row vector represents a particular word or term and the columns represent the documents. Thus, each row captures statistical distributional information of a word across documents and each column represents distributional information of the words in a document, i.e., co-occurrence information among words within a single document. Furthermore, this statistical distributional information of a word/term across documents and of word/term co-occurrence within a document is captured in an elegant algebraic format of a term-by-document matrix. This representation then enables the use of algebraic operations, an important attraction of this representation, serving different purposes. For instance, multiplying the term-by-document matrix by its transpose results in a term-by-term co-occurrence matrix where each cell provides a quantitative summary of how two words/terms co-occur with each other across all documents in the collection.

To construct the term-by-document matrix, many different approaches may be used. The original information retrieval approach is a frequency approach where words are counted over documents. Thus, a given $cell_{ij}$ of the matrix represents the number of times that $word_i$ appears in $document_j$. Because not all words occur in all documents, the row vectors are necessarily sparse. The similarity between words can be computed based on the similarity of their vectors, using metrics like dot product or cosine, i.e., normalized dot-product, to eliminate bias toward longer documents. Furthermore these word vectors have the property that they may be added together, allowing one to obtain a vector of a chunk of text, e.g., a phrase or sentence or paragraph, by adding the vectors for words in the text. This additive process to obtain the meaning of a larger text from its individual words is a simple instantiation of the meaning compositionality principle (the simple co-occurrence in the same text is, in this case, the rule-to-combine or "relationship" specified in the definition of the principle of compositionality mentioned earlier). The resulting vectors can be compared with any other vector using the same vector similarity metrics.

Although this basic term-by-document structure is shared by all members of the vector space family, variations are wide ranging. For example, words can be preprocessed before computing their frequencies using techniques like stemming (Rus, Banjade & Lintean, 2014). Rus, Banjade, and Lintean (2014) discuss in detail the implication of preprocessing steps on the performance of semantic processing approaches and show that such preprocessing steps have a much more significant impact than recognized by the research community (it is important to note that they identified and considered 1,152 combinations of preprocessing steps in their study). Raw counts may also be transformed in a variety of ways using information about the cell, its row, or its column. For example, the presence of a word in a document is typically more important than the number of times the word appeared, leading to word frequency normalization like log scaling of cell counts, and words that appear in many documents are less diagnostic of meaning and so can be down-weighted using row weighting schemes like inverse document frequency; similarly approaches based on information theory can be used to weight cell counts based on word (row) or column (document) distributions (Church & Hanks, 1989; Dumais, 1991; Manning & Schötze, 1999).

The set-based similarity approaches that we mentioned earlier can be regarded as using an underlying vector representation with binary weights: the word is present, corresponding to a value of 1, or not, corresponding to a value of 0. The dot product between such binary vectors corresponding to two texts results in the cardinality/size of the intersection of the underlying sets of words, i.e., the number of common words in the two texts. When using raw frequency as the weights, the result is a similar vector representation, which regards the texts as bag of words (multiple occurrences of the same word are accounted for) as opposed to sets of words (multiple occurrences of the same word only count once). In fact, we can think of a vector representation for individual words even if an explicit one is not derived from a corpus. Each word can be thought of as having a 1-of-N representation, where N is the size of the vector, i.e., the number of entries or dimensionality (and equal to the vocabulary size of the two texts), and only one cell or entry that corresponds to dimension of the target word has a weight of 1 (all other entries being zero).

Finally, some approaches transform the matrix into another matrix using matrix factorization techniques. In particular, the use of singular value decomposition to factorize the matrix is strongly associated with the vector space approach known as latent semantic analysis (LSA; Landuaer, Foltz & Laham, 1998; Landauer, McNamara, Dennis & Kintsch, 2007). The computational advantage of LSA is that it represents that meaning of words using a reduced dimensionality space (300–500 dimensions) leading to fast computations of similarity scores, e.g., based on cosine similarity. A study comparing the effect of various local and global weighting schemes in conjunction with LSA was described by Lintean, Moldovan, Rus, and McNamara (2010).

Although the basic approach outlined previously seems very sensible from an information retrieval perspective focused on documents, from a generic text similarity perspective, aspects of this approach are

somewhat arbitrary. For example, documents are not defined to have any particular length, leading to questions about whether there is an optimum length for documents or whether it matters if document lengths across a matrix are highly variable. Likewise, it is useful to have words as the unit of co-occurrence analysis and representation but in many applications of natural language processing sequences of words called n-grams have been used with great success. Considerations such as these have led to proposals for abstracting from the word by document matrix to a feature by context matrix (Olney, 2009). In these matrices, the concern is counting a particular feature within a given context, where both context and feature can be arbitrarily defined. For example, a square matrix could be defined with rows and columns equal to the number of words in the collection, so a $cell_{ij}$ might represent the number of times $word_i$ occurs after $word_j$. Likewise a row could be a word and $cell_{ij}$ might represent the number of times $word_i$ is the syntactic head of $word_j$.

Taking a feature by context perspective is useful when considering approaches where there is no document or the document is part of a larger interconnected structure. For example, the Correlated Occurrence Analogue to Lexical Semantics model (COALS; Rohde, Gonnerman & Plaut, 2005; Olney, Dale & D'Mello, 2012) implements a sliding window strategy. Similar to the previous example, the resulting matrix is square, but the cell counts are based on a symmetric window around the target word. The word at the center of that window corresponds to the row of the matrix, and the incremented columns are the other words within the window; however, the increment is inversely proportional to the distance between them and the target word. The matrix is normalized using phi correlation and then optionally transformed using singular value decomposition.

Structured texts, like Wikipedia, have inspired structure-specific approaches defining features and contexts. Explicit Semantic Analysis (ESA) uses the article structure of Wikipedia to determine contexts, and thus a $cell^{ij}$ represents the number of times $word_i$ occurs in $article_j$. ESA further weights the cells using log weighting of word counts and inverse document frequency. Another approach that uses the structure of Wikipedia to define both features and contexts is Wikipedia Link Measure (WLM; Milne & Witten, 2008). WLM uses articles as features and links to them from other Wikipedia pages as contexts. To compare the similarity of two words, WLM associates them with articles and then defines a metric over the shared links to those pages (in-links) and shared links from those pages (out-links) to determine similarity. Although the original definition of WLM is more set theoretic based on the graph structure of links, it trivially maps to a vector space approach where links are columns in a matrix (Olney et al., 2012).

One other word meaning representation worth noting is Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan, 2003). LDA is a probabilistic, unsupervised method that models documents as distributions over topics and topics as distributions over words in the vocabulary. Each word has a certain contribution to a topic. That is, each word can be represented as a vector whose dimensionality equals the number of latent topics and the weights along each dimension correspond to the word contributions to each topic as derived by the LDA method. Based on these distributions and contributions word-to-word semantic similarity and text-to-text semantic similarity measures can be defined (Rus, Niraula & Banjade, 2013). It should be noted that LDA has a conceptual advantage over LSA because LDA explicitly represents different meanings of words, i.e., it captures polysemy. In LDA, each topic is a set of words that together define a meaning or concept, which corresponds to a specific sense of the various words sharing this meaning. Thus in LDA, each topic a word belongs to can be regarded as one of its senses. On the other hand, LSA has a unique representation for a word. That is, all meanings of a word are represented by the same LSA vector making it impossible to distinguish among the various senses of the word directly from the representation. Some argue that the LSA vector represents that dominant meaning of a word while others believe the LSA vector represents an average of all the meanings of the word. The bottom line is that there is a lack of explicit account for the various senses of a word in LSA.

Yet another major trend in the area of distributional vector-based representations, is the new category of representations derived using deep neural networks. A typical example in this category is the word2vec

representation developed by Mikolov and colleagues (2013). They trained a recursive neural network with local context (continuous n-gram or skip-gram) using a process in which the input word vectors are recursively adjusted until the target context words are accurately predicted given the input word. These representations were shown to capture syntactic and lexical semantic regularities, have superior compositionality properties, and enable precise analogical reasoning using simple vector algebra (Mikolov, Chen, Corrado & Dean, 2013; Pennington, Socher & Manning, 2014).

A Categorization of Compositional based Semantic Similarity Approaches

Once a word meaning representation is being settled upon, the meaning of larger texts can be derived compositionally in many different ways. We distinguish among the following major categories of compositional approaches.

Simple Additive Approaches

Simple additive approaches generate an overall semantic similarity score by applying simple operations, such as addition and average matching, to word level representations. Two examples of such simple additive approaches are the vector addition and word-average matching addition (Fernando & Stevenson, 2008). In the vector addition approach, an overall vector for each text is computed by simply adding up the vector representations of individual words. Then a vector similarity metric is applied on the resulting vectors of the two target texts. In the word-average matching addition approach, for each word in one text an average semantic similarity score is computed with all the words in the other text. To compute this average similarity score for a single word, the word is matched or paired with every single word in the other text and a similarity metric is computed between the corresponding vector representations of the paired words. The average of all those word-to-word similarity scores is then taken.

Alignment-Based Approaches

In this category, the approaches have a distinguished feature of aligning first the words in one text to at most one word in the other text. From a student answer assessment point of view, performing the alignment first has the great advantage of detecting which main concepts the student articulated and which ones are missing from the response. Detecting the articulated and missing concepts, in turn, enables automated generation of feedback and informs the dialogue/interaction planning component of the underlying ITS.

The alignment can be based on simple lexical matching (using the words themselves) in which case we end up with a simple lexical overlap score, i.e., the number of common words between the two texts. This simple lexical overlap score is computed after applying the various preprocessing steps chosen by the designer such as lemmatization and can be normalized by the average length or the maximum length or some other normalizing factor (see Lintean, Moldovan, Rus & McNamara, 2010; Rus, Banjade & Lintean, 2014). The two texts can be treated as sets or as bags. In the latter case, words that match can be weighted by their frequencies in the two texts to account for multiple occurrences.

A more advanced approach, called lexical similarity alignment, relies on word-to-word (w2w), or lexical, semantic similarity measures to find the best alignment between a word in one text and at most one word in the other text. We distinguish between two types of lexical similarity alignment approaches: *greedy lexical similarity alignment* and *optimal lexical similarity alignment* that only relies on lexical similarities between words but not on their relationship (e.g., syntactic and/or deeper semantic relations). The greedy approach simply aligns a word in one text with the word in the other text that leads to the highest w2w similarity score according to some w2w similarity measures.

The optimum assignment between words in one text, T1, and words in another text, T2, can be posed as a combinatorial optimization problem. That is, the goal is to find a permutation π for which $\sum_{k=1}^{n} \blacksquare \Theta(v_i, w_{\pi(i)})$ is maximum, where Θ denotes any word-to-word similarity measure, and v and w are words from the texts T1 and T2, respectively. This formulation of the T2T similarity problem is in fact the famous job assignment problem for which an algorithm, the Kuhn-Munkres or Hungarian method (Kuhn, 1955), has been proposed and which can find a solution in polynomial time.

The assignment problem only focuses on optimally matching words in one sentence S to words in the other sentence T based only on how the words in S match the words in T. As briefly mentioned, it does not account for interdependencies among words in S or among words in T. A solution that simultaneously accounts for such inter-dependencies, thus capturing the context of each word in their corresponding sentences, has been proposed by Lintean and Rus (2015).

The *optimal lexico-relational alignment* method aims at finding an optimal global assignment of words in one sentence (e.g., a student response) to words in the other sentence (e.g., the expert answer) based on their w2w similarity, while simultaneously maximizing the match between the syntactic dependencies. Accounting for the syntactic dependencies among words is the primary advantage of the quadratic assignment problem (QAP; Koopmans-Beckmann, 1957) formulation versus the job-assignment formulation of the student response assessment task (Rus & Lintean, 2012).

The formulation of the QAP problem for textual semantic similarity proposed by Lintean and Rus (2015) is to maximize the objective function QAP (see below), where matrix F and D describe dependencies between words in one sentence and the other, respectively, while B captures the w2w similarity between words across the two texts. Also, they weighted each term resulting in the following formulation:

$$max\ QAP(F, D, B) = \alpha \sum_{i=1}^{n} \sum_{j=1}^{n} f_{i,j} d_{\pi(i)\pi(j)} + (1-\alpha) \sum_{i=1}^{n} b_{i,\pi(i)}$$

The $f_{i,j}$ term quantifies the (syntactic or semantic or of other nature) relation between words i and j in text A which are mapped to words $\pi(i)$ and $\pi(j)$ in text B, respectively. The distance $d_{\pi(i)\pi(j)}$ quantifies the semantic relation between words $\pi(i)$ and $\pi(j)$. For words i and j that have a direct syntactic dependency relation, i.e., an explicit syntactic relation among two words such as subject or direct object, the "flow" $f_{i,j}$ is set to 1 (0 – if not direct relation). Similarly, the distance $d_{\pi(i)\pi(j)}$ between words $\pi(i)$ and $\pi(j)$ is set to 1 in case there is a direct dependency relation among them and 0 otherwise.

A brute force solution to the QAP problem, which would generate all possible mappings from words in a sentence to words in the other sentence, is not feasible because the solution space is too large. For example, when considering all possible pairings of words between sentence A, of size n, and sentence B of size m, where n < m, and there are no limitations on the type of pairings that can be made, there are $m!/(m-n)!$ possible solutions. For sentences of average size $n = m = 20$ words, there are 2.4 * 10^{18} possible pairings. An branch-and-bound algorithm was proposed by Lintean and Rus (2015) to efficiently explore the solution space in search for the optimal solution.

Interpretable alignment based approaches perform first an alignment between words in one text versus words in the other texts while at the same time identifying semantic labels for aligned words (Banjade, Maharjan, Gautam, and Rus, 2016). The advantage of adding semantic relationships between the aligned words is that an explanation for the alignment can be provided based on these w2w semantic relationships. The set of semantic relationships used were proposed as part of the interpretable Semantic Textual Similarity (iSTS) task (Agirre et al., 2016), organized by SemEval – the leading semantic evaluation forum, and includes: EQUI (semantically equivalent), OPPO (opposite in meaning), SPE (one chunk is more specific

than other), SIMI (similar meanings, but not EQUI, OPPO, SPE), REL (related meanings, but not SIMI, EQUI, OPPO, SPE), and NOALI (has no corresponding chunk in the other sentence). It should be noted that we presented alignment based methods focusing on words. Other units of analysis can be used such as chunks (Stefanescu, Banjade & Rus, 2014a).

Resources

Aside from a quantitative and qualitative outcome for T2T tasks, there recently has been a push to offer an explanation of the final T2T outcome. To this end, several resources and efforts have been reported as explained below. Additionally, we list some other relevant resources that have been developed and released publicly in order to foster research in this area.

The SEMILAR Corpus. Rus and colleagues (2012) developed the SEMILAR corpus, which is the richest in terms of annotations as besides holistic judgments of paraphrase they provide several word level similarity and alignment judgments. The corpus includes a total of 12,560 expert-annotated relations for a greedy word-matching procedure and 15,692 relations for an optimal alignment procedure.

The Student Response Analysis (SRA) Corpus. The SRA corpus (Dzikovska et al., 2013) consists of student answer-expert answer pairs collected from two ITSs. Both student answers and expert answers were related to specific tutorial questions from different science domains. There are 56 questions and 3,000 student answers from the so-called BEETLE corpus as well as 197 assessment questions and 10,000 answers from the ScientsBank corpus. These pairs were annotated using a combination of heuristics and manual annotation. They used a 5-way annotation as opposed to the typical 2-way annotation.

The User Language Paraphrase Corpus (ULPC; McCarthy and McNamara 2008). ULPC contains pairs of target sentence/student response texts. The student responses were collected from experiments with the ITS iSTART. Students were shown individual sentences collected from biology textbooks and asked to paraphrase them. These pairs have been evaluated by expert human raters along 10 dimensions of paraphrase characteristics. The "paraphrase quality bin" dimension measures the paraphrase quality between the target-sentence and the student response on a binary scale. From a total of 1,998 pairs, 1,436 (71%) were classified by experts as being paraphrases. A quarter of the corpus is set aside as test data. The average words per sentence is 15.

DeepTutor Anaphora Resolution Annotated (DARE): The DARE corpus (Niraula, Rus, Banjade, Stefanescu, Baggett, and Morgan, 2014) is an annotated data set focusing on pronoun resolution in tutorial dialogue. Although data sets for general purpose anaphora resolution exist, they are not suitable for dialogue-based ITSs, which is the reason the DARE corpus was created. The DARE corpus consists of 1,000 annotated pronoun instances collected from conversations between high school students and the ITS DeepTutor. The data set is publicly available.

DeepTutor Tutorial Dialogue Negation Annotated (DT-Neg; Banjade and Rus, 2016) Corpus. Negation is found more frequently in dialogue than typical written texts, e.g., literary texts. Furthermore, the scope and focus of negation depends on context in dialogues more so than in other forms of texts. Existing negation data sets have focused on non-dialogue texts such as literary texts where the scope and focus of negation is normally present within the same sentence where the negation cue is located and therefore are not the most appropriate to inform the development of negation handling algorithms for dialogue-based systems. The DT-Neg corpus contains texts extracted from tutorial dialogues where students interacted with the ITS to solve conceptual physics problems. The DT-Neg corpus contains 1,088 annotated negations in student responses with scope and focus marked based on the context of the dialogue.

DeepTutor Student Answer Grading Context-aware Annotated (DT-Grade; Banjade, Maharjan, Niraula, Gautam, Samei & Rus, 2016) *Corpus*. The DT-Grade corpus consists of short constructed answers extracted from tutorial dialogues between students and the DeepTutor ITS and annotated for their correctness in the given context and whether the contextual information was useful. The data set contains 900 answers of which about 25% required contextual information to properly interpret.

Semantic Similarity Toolkit (*SEMILAR;* Rus, Lintean, Banjade, Niraula & Stefanescu, 2013). The SEMILAR software package offers users (researchers, practitioners, and developers) easy access to fully-implemented semantic similarity methods in one place through both a graphical user interface (GUI)-based interface and a software library. Besides productivity advantages, SEMILAR provides a framework for the systematic comparison of various semantic similarity methods. The automated methods offered by SEMILAR range from simple lexical overlap methods to methods that rely on w2w similarity metrics to more sophisticated methods that rely on fully unsupervised methods to derive the meaning of words and sentences such as LSA and LDA to kernel-based methods for assessing similarity.

Conclusions and Recommendations for Future Research

Given the importance of student-generated open responses in educational context, which, we argue, are the only assessment modality that leads to true assessment because are the only assessment modality that reveals students' true mental model, future educational technologies including the Generalized Intelligent Framework for Tutoring (GIFT; Sottilare, Brawner, Goldberg & Holden, 2012; Sottilare, Brawner, Sinatra & Johnston) should include open-ended assessment items and corresponding facilities that enable the automated assessment of such open-ended student responses.

True assessment is necessary to infer an accurate picture of students' mastery level, which in turn is paramount for triggering appropriate feedback and instructional strategies and ultimately the effective and efficiency of the underlying educational technology. Considering the early stage nature of the assessment module in the educational processing pipeline and therefore the positive or negative cascading effect it may have on the downstream modules (learner model, feedback, strategies, and outcome, e.g., learning) the importance of automated assessment of open-ended learner responses cannot be overstated.

Assessing students' open-ended responses is complex and requires a multitude of factors to be considered, as illustrated in this contribution. However, this complexity is surmountable and there has been tremendous progress in terms of advanced methods and resources that have been developed and publicly released.

Acknowledgments

The research was supported in part by the Institute for Education Sciences under award R305A100875, under Cooperative Agreement W911NF-12-2-003 between U.S. Army Research Laboratory (ARL) and the University of Memphis, Office of Naval Research (N00014-00-1-0600, N00014-15-P-1184; N00014-12-C-0643; N00014-16-C-3027) and the National Science Foundation Data Infrastructure Building Blocks program (ACI-1443068). Any opinions, findings, and conclusions expressed are solely the authors'.

References

Agirre, E., Gonzalez-Agirre, A., Lopez-Gazpio, I., Maritxalar, M., Rigau, G., and Uria, L. (2016). Semeval-2016 task 2: Interpretable semantic textual similarity. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June, 2016.Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... Schneider, N. (2013, August). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse (pp. 178–186)*. Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W13-2322

Banjade, R., Niraula, N., and Rus, V. (2016). Towards Intra- and Inter-Sentential Negation Scope and Focus Detection in Tutorial Dialogue. *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2016)*, Key Largo Resort, Key Largo, Florida, USA May 16–18, 2016.

Banjade, R., Maharjan, N., Gautam, D., and Rus, V. (2016). DTSim at SemEval-2016 Task 1: Semantic Similarity Model Including Multi-Level Alignment and Vector-Based Compositional Semantics, In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies will be held in San Diego, California, June 12–17, 2016.

Banjade, R. and Rus, V. (2016). DT-Neg: Tutorial Dialogues Annotated for Negation Scope and Focus in Context, *Proceedings of The 10th International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia, May 23–28, 2016.

Banjade, R., Maharjan, N., Niraula, N.B., Gautam, D., Samei, B., and Rus, V. (2016). Evaluation Dataset (DT-Grade) and Word Weighting Approach towards Constructed Short Answers Assessment in Tutorial Dialogue Context, *The 11th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL 2016)*, The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies will be held in San Diego, California, June 12 to June 17, 2016.

Bender, E. M., Flickinger, D., Oepen, S., Packard, W. & Copestake, A. (2015, April). Layers of interpretation: On grammar and compositionality. *In Proceedings of the 11th international conference on Computational Semantics* (pp. 239–249). London, UK: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W15-0128.

Beigman-Klebanov, B., Madnani, N., Burstein, J. C. & Sumasundaran, S. (2014). Content importance models for scoring writing from sources. *Proceedings of the 52 Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, pp. 247–252.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003). Latent Dirichlet Allocation, *The Journal of Machine Learning Research*, 3, 993–1022.

Bos, J. (2015, May). Open-domain semantic parsing with boxer. In *Proceedings of the 20th nordic conference of computational linguistics (NODALIDA 2015)* (pp. 301–304). Vilnius, Lithuania: Linko¨ping University Electronic Press, Sweden. Retrieved from http://www.aclweb.org/anthology/W15-1841

Burstein, J., Chodorow, M. & Leacock, C. (2004). Automated essay evaluation: The Criterion Online writing service. *AI Magazine*, 25(3), 27–36.

Church, K. W. & Hanks, P. (1989, June). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting of the association for computational linguistics* (pp. 76–83). Vancouver, British Columbia, Canada: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P89-1010 doi:10.3115/981623.981633.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods*, Instruments & Computers, 23(2), 229–236. Retrieved from http://dx.doi.org/10.3758/BF03203370 doi: 10.3758/BF03203370.

Dzikovska, M.O., Nielsen, R.D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I. and Dang, H.T. (2013). SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, 7th International Workshop on Semantic Evaluation (SemEval 2013) Atlanta, Georgia, USA, June 13-14, 2013. Published by the Association for Computational Linguistics.

Fernando, S. & Stevenson, M. (2008). A semantic similarity approach to paraphrase detection. In *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, pages 45–52.

Foltz, P. W., Gilliam, S. & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments*, 8(2), pp. 111–129.

Foltz, P. W., Laham, D. & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Education Journal of Computer Enhanced Learning*, 1, (2).

Foltz, P. W., Streeter, L. A., Lochbaum, K. E. & Landauer, T. K (2013). Implementation and applications of the Intelligent Essay Assessor. *Handbook of Automated Essay Evaluation*, M. Shermis & J. Burstein, (Eds.). Pp. 68-88. New York: Routledge.

Higgins, D., Brew, C., Hellman, M., Ziai, R., Chen, L., Cahill, A., … Blackmore, J. (2014) Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. arXiv: 1404.0801, v2.

Koopmans, T. C. and Beckmann, M. 1957 Assignment Problems and the Location of Economic Activities. *Econometrica*, volume 25(1), pages 53-76. The Econometric Society.

Kuhn, H.W. (1955). The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, volume 2:83–97.

Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2&3), 259–284.

Landauer, T. K, Laham, D. & Foltz, P. W. (2001). Automated essay scoring. *IEEE Intelligent Systems*. September/October.

Landauer, T. K., McNamara, D. S., Dennis, S. & Kintsch, W. (Eds.). (2013). *Handbook of latent semantic analysis*. New York: Routledge.

Lenat, D.B. (1995). Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*. 38

Liang, P., Jordan, M. I. & Klein, D. (2013). Learning dependency-based compositional semantics. *Computational Linguistics*, 39(2), 389–446.

Lintean, M., Moldovan, C., Rus, V. & McNamara, D. (2010). The Role of Local and Global Weighting in Assessing The Semantic Similarity Of Texts using Latent Semantic Analysis. *Proceedings of the 23st International Florida Artificial Intelligence Research Society Conference*, Daytona Beach, FL.

Lintean, M. & Rus, V. (2015). An Optimal Quadratic Approach to Monolingual Paraphrase Alignment, *The 20th Nordic Conference on Computational Linguistics (NODALIDA 2015)*, Lithuania, Vilnius, 11–13 May 2015.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Milne, D. & Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy* (pp. 25–30). Chicago: AAAI Press.

Moldovan, D.I. and Rus, V. (2001). Logic Form Transformation of WordNet and its Applicability to Question Answering, *Proceedings of the ACL 2001 Conference*, July 2001,Toulouse, France.

Niraula, N., Rus, V. & Stefanescu, D. (2013). DARE: Deep Anaphora Resolution Engine. *Proceedings of The 6th International Conference on Educational Data Mining (EDM 2013)*, July 6–9, Memphis, TN.

Niraula, N., Rus, V., Banjade, R., Stefanescu, D., Baggett, W. and Morgan, B. (2014). The DARE Corpus: A Resource for Anaphora Resolution in Dialogue Based Intelligent Tutoring Systems, *The 9th International Conference on Language Resources and Evaluation*, Reykjavik, May 26–31, 2014, Iceland.

Olney, A. M. (2009). Generalizing Latent Semantic Analysis. In 2009 *IEEE International Conference on Semantic Computing* (pp. 40–46). https://doi.org/10.1109/ICSC.2009.89.

Olney, A. M., Dale, R. & D'Mello, S. K. (2012). The World Within Wikipedia: An Ecology of Mind. *Information*, 3(2), 229–255. https://doi.org/10.3390/info3020229.

Pennington, J., Socher, R., Manning, C.D. (2014). Glove: Global vectors for word representation. Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014) 12, 1532–1543.

Perelman, L. C. (2013). Critique of Mark D. Shermis & Ben Hamner, Contrasting State-of-the-Art Automated Scoring of Essays: Analysis", *The Journal of Writing Assessment*, Volume 6, Issue 1: 2013.

Rohde, D., Gonnerman, L. & Plaut, D. (2005). An improved model of semantic similarity based on lexical co-occurrence. Retrieved from http://tedlab.mit.edu/dr/Papers/RohdeGonnermanPlaut-COALS.pdf.

Rosé, C. P., Gaydos, A., Hall, B.S., Roque, A. & VanLehn, K. (2003). Overcoming the Knowledge Engineering Bottleneck for Understanding Student Language Input. In H. U. Hoppe, F. Verdejo and J. Kay (Eds.), *Artificial Intelligence in Education.*

Rus, V. (2002). *Logic Form For WordNet Glosses and Application to Question Answering*, Computer Science Department, School of Engineering, Southern Methodist University, PhD Thesis, May 2002, Dallas, Texas.

Rus, V. & Graesser, A.C. (2006). Deeper Natural Language Processing for Evaluating Student Answers in Intelligent Tutoring Systems, *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*.

Rus, V., Niraula, N. & Banjade, R. (2013). Similarity Measures based on Latent Dirichlet Allocation. *The 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*, March 24–30, Samos, Greece.

Rus, V., Lintean, M., Banjade, R., Niraula, N. & Stefanescu, D. (2013). SEMILAR: The Semantic Similarity Toolkit. *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, August 4–9, Sofia, Bulgaria.

Rus, V., D'Mello, S., Hu, X. & Graesser, A.C. (2013). Recent Advances in Conversational Intelligent Tutoring Systems, *AI Magazine*, 34,(3):42–54.

Rus, V., Banjade, R., and Lintean, M. (2014). On Paraphrase Identification Corpora, *The 9th International Conference on Language Resources and Evaluation*, Reykjavik, May 26–31, 2014, Iceland.

Salton, G., Wong, A. & Yang, C. S. (1975). A vector space model for automatic indexing. Commun. *ACM*, 18(11), 613–620. doi: http://doi.acm.org/10.1145/361219.361220.

Shermis, M.D., and Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and Future Directions*. New York: Routledge.

Shermis, M. and Hamner. B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. In presented at *Annual Meeting of the National Council on Measurement in Education*, Vancouver, Canada, April.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76.

Sottilare, R.A., Brawner, K.W., Goldberg, B.S. & Holden, H.K. (2012). The Generalized Intelligent Framework for Tutoring (GIFT). Concept paper released as part of GIFT software documentation. Orlando, FL: U.S. Army Research Laboratory – Human Research & Engineering Directorate (ARL-HRED). Retrieved from: https://gifttutoring.org/attachments/152/GIFTDescription_0.pdf

Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017. DOI: 10.13140/RG.2.2.12941.54244.

Ştefănescu, D., Banjade, R., Rus, V. (2014). Latent Semantic Analysis Models on Wikipedia and TASA. *The 9th International Conference on Language Resources and Evaluation*, May 26-31, 2014, Reykjavik, Iceland.

Ştefănescu, D., Banjade, R., Rus, V. (2014a). A Sentence Similarity Method based on Parsing and Information Content. In *Proceedings of 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014)*, April 6–12, Kathmandu, Nepal.

Thurlow, M., Hermann, A. & Foltz, P. W. (2010). Preparing MSA Science Items for Artificial Intelligence (AI) Scoring. Talk presented at the Maryland Assessment Group Conference. Ocean City, November.

VanLehn, K., Jordan, P., Rosé, C. P., et al. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S. A. Cerri, G. Gouarderes & F. Paraguacu (Eds.), *Intelligent Tutoring Systems: 6th International Conference* (pp. 158–167).

Walstad, W.B. & W.E. Becker (1994). Achievement differences on multiple-choice and essay tests in economics. *Proceedings of the American Economic Association*, 193–196.

Williamson, D., Xi, X. & Breyer, J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31, 2–13.

Williamson, D. M., Bennett, R., Lazer, S., Bernstein, J., Foltz, P. W., … Sweeney, K. (2010, June). Automated Scoring for the Assessment of Common Core Standards. Retrieved July 1 2014 from http://www.ets.org/s/commonassessments/pdf/AutomatedScoringAssessCommonCoreStandards.pdf

Zinn, C., Moore, J.D., Core, M.G., Varges, S. & Porayska-Pomsta, K. (2003). The BE&E Tutorial Learning Environment (BEETLE). System demonstration in *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck)*.

CHAPTER 14 – Using Process Data for Assessment in Intelligent Tutoring Systems: A Cognitive Psychologist, Psychometrician, and Computer Scientist Perspective

Samuel Greiff[1], Dragan Gasevic[2], and Alina A. von Davier[3]
University of Luxembourg[1], University of Edinburgh[2], ACTNext by ACT, Inc[3]

Introduction

Process data offer potentially rich information on how test-takers navigate through complex assessment and learning environments such as intelligent tutoring systems (ITSs). However, the actual exploitation of these data strings for facilitation of assessment and learning has proven considerably more difficult than initially anticipated and has often lacked the interdisciplinary efforts needed for deriving a comprehensive perspective. In this chapter, a cognitive psychologist, a psychometrician, and a computer scientist elaborate on the most prominent challenges they deem important when exploiting the large amounts of computer-generated process data and when trying to reveal the fuzzy relations therein. In this, three different perspectives are presented that, in an effort to contribute to an interdisciplinary discussion, are then investigated with regard to their potential of convergence toward the common goal of making ITSs strong facilitators of learning and assessment.

Over the last decades, tools aimed at measuring and enhancing a variety of skills and competencies have undergone tremendous innovation, most visibly associated with a comprehensive shift to computer-based assessment and computer-supported learning tools. One class of these sophisticated learning tools are ITSs. ITSs are complex computer-simulated environments that help "the student master deep knowledge/skills by implementing powerful intelligent algorithms that adapt to the learner at a fine-grained level and that instantiate complex principles of learning" (Graesser, Hu, Nye & Sottilare, 2016, p. 60) and in which a number of tasks and problems need to be solved throughout the process. Depending on the set of actions a person performs when working on an ITS, feedback and specifically targeted support is offered by the system. Importantly, the overall course of information and support offered by the system might vary depending on the individual proficiency levels. Thus, along the way, several learning experiences, most of them individually tailored to the learner, are offered that are aimed at enhancing the target skills.

For instance, one typical ITS is *Operation ARIES* (Millis et al., 2011). Operation ARIES sets up an environment that targets a student population within a context of an alien invasion of the earth. Within several missions, test-takers need to solve tasks that all require scientific reasoning and evaluation of scientifically rigorous studies. An automated tutor helps test-takers and provides support to them with the aim of facilitating students' qualitative science inquiry skills. However, ITS are not limited to student populations in K–12 or college. There are a number of ITS that mirror complex real-world scenarios and are used for training purposes in the military and business communities.

The overarching purposes of ITSs are 1) to elicit skills and competencies, such as planning skills, teamwork, or scientific inquiry, that are relevant in academic and real-world contexts; and 2) to make use of automated feedback and support within a computer-based environment to enhance and facilitate these skills. Obviously, the implicit hope behind this rationale is that learning that occurs within ITSs will transfer to better performance outside of the ITS as well. This is a highly challenging task and it is no surprise that developing, introducing, and using a scientifically sound and valid ITS involves substantial effort and usually requires interdisciplinary cooperation that combines expertise from a number of different areas.

This chapter illustrates the added value of an interdisciplinary understanding of the development and proper application of ITSs that includes views from three crucial areas: cognitive psychology, psychometrics, and computer science. Whereas cognitive psychology can contribute to the understanding of what happens in the human brain and which cognitive processes are involved when working on ITSs, psychometrics can show ways of how scientifically sound indicators can be derived out of the seemingly endless data streams provided by ITSs. Computer science needs to set constraints and explore possibilities of what is technically possible, how the content of an ITS can be put into action and how it connects to the field of computer science. It is only in their combination and interplay that interdisciplinary efforts can evolve into palpable outputs, that is, into useful ITSs of high quality and validity. In this chapter, a representative from each of the three prominent fields mentioned above (cognitive psychology, psychometrics, and computer science) discusses from their personal perspective the challenges in developing an ITS. This is not the ultimate word but more like a starting point for a collaborative discourse. At the end, we offer some suggestions on how the Generalized Intelligent Framework for Tutoring (GIFT) might accommodate the expertise from multiple fields.

The Cognitive Psychologist's View

"Targeting the cognitively relevant processes and adequately mapping them to theories of the human mind."

ITSs produce data. A lot of data. When a person interacts with an ITS, there is a continuous stream of data that emerges, including active interventions, answers to questions, inquiries, timestamps for each action, overall performance, and so forth. To researchers, and in fact to anybody who wants to understand what happens in the course of an ITS, this data stream is tempting but it is also dangerous. It is dangerous for at least two reasons. The first is that one quickly gets lost in the almost infinite amount of data and that it is extremely difficult to actually extract the often fuzzy relations between learning, performance, and single actions. The second and even more worrying is that when facing such amounts of data and when discovering some empirical relations, it is all too easy to forget about what behaviors actually stand for and what kind of underlying process they indicate (or do not indicate), even though this is fundamental for any valid interpretation.

More specifically, for cognitive psychology, it is not so much the overt behavior that is interesting and relevant, but rather the underlying cognitive processes that are of interest. For instance, it is not intrinsically interesting that Cindy shows a steep learning curve and solves most of the ITS tasks correctly after working with the ITS for a while, whereas Ben does not. The interesting question is whether these differences can be mapped onto some differences in an underlying cognitive skill or competency, such as scientific reasoning, critiques, and inquiry, the target skills in "Operation ARIES".

A theory from cognitive science on the underlying processes is needed as a firm starting point. For example, in the field of individual problem solving, several processes are distinguished over the course of problem solving. In the Programme for International Student Assessment (PISA; cf. OECD, 2014 for details), there are four theoretical processes in problem solving: exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting. Consequently, different tasks are developed that target the different processes to varying extents. In a similar way, in the field of science inquiry, several aspects of the overall skill (or set of skills) are separated. For instance, Gobert, Sao Pedro, Baker, Toto, and Montalvo (2012) distinguish between hypothesizing, experimenting, interpreting data, and communicating, with three out of these four separated into further subskills.

These theoretically motivated conceptions about the underlying cognitive processes could serve as starting point of what the actual target dimensions are in ITSs and beyond, as compared to specific behaviors that

are unconnected to an underlying theory. However, there is usually no one-to-one mapping between specific behaviors and the cognitive processes associated with them and the reasons for this are manifold with one of them being the inherent difficulty of mapping specific behaviors onto cognitive processes. This is not surprising given that even high-resolution imaging methods of the brain allow only to a limited extent for straightforward relations between mental performance and activation of brain areas (but see Haier, 2016, for a worthwhile read on what neuroscience can tell us about intelligence).

Put differently, there is no isomorphic mapping between a specific behavior and a specific cognitive process. Time-on-task (TOT), as an example, has been targeted in a large body of research on cognitive performance and has frequently caught researchers' attention. However, measures of TOT can mean very different things; a short TOT can be indicative both of immediate knowledge of the answer and of low motivation to engage with the task. A long TOT, on the other hand, can be indicative of in-depth cognitive processing as well as of slow reading time when the task involves some written instruction. In fact, we know surprisingly little on how the connection between specific behavioral actions and patterns on the one hand and cognitive processes on the other hand can be firmly established. However, establishing this connection is crucial and the only reasonable starting point for this is sound cognitive theory.

An example of how a general framework on cognition can serve as starting point is provided by the work of Goldhammer et al. (2014). Within a dual-processing theoretical framework (e.g., Schneider & Chein, 2003), the authors derive differential expectations of how TOT relates to performance measures in two different types of tasks: problem solving and reading. The results of Goldhammer and colleagues confirm that individual differences in TOT are positively related to overall performance differences in problem solving as a cognitive task that requires large amounts of control, whereas there is a negative relation between TOT and reading performance in some reading materials that require largely automatized cognitive processing with low levels of complexity. Thus, depending on the level of automatization, either short or long TOTs indicate the adequate allocation of resources to a task.

In a similar way, depending on the level of control (i.e., automatic vs. controlled processing), interindividual differences in TOT when working on ITSs might be an important marker of the type of the underlying processing going on. Even further, intra-individual differences in TOT (i.e., differences over time within one individual), for instance, at the beginning and at the end of working on an ITS, might be indicative of a change in resource allocation and worth analyzing. That is, a person working on an ITS might learn how to appropriately allocate time resources depending on the level of control needed. If so, the intra-individual profile of TOT could be an indicator of learning (for an example see Forsyth, Graesser, Pavlik, Millis & Samei, 2014).

However, the core point from the perspective of a cognitive psychologist is that some kind of theoretical embedding is required to map specific behaviors onto cognitive processes and integrating them into a theory on the human mind. Goldhammer et al. (2014) employ a rather broad and general theory that serves as an example of how such a theoretical connection can be readily (and even rather easily) achieved. Obviously, this is not a mere theoretical effort, but requires empirical validation of the theoretical assumptions in a second step. This research-driven effort may sometimes conflict with the pragmatism needed when developing ITS for particular practical applications. However, in the long run, the validity and the utility of ITSs will greatly benefit from explicated efforts along theory guided design and systematic empirical methodologies.

The Psychometrician's View

"Objectively measuring the skills, abilities, and educational achievement relevant in ITSs with statistical tools embedded in measurement theory."

Technology is a powerful foundation for the assessment of complex skills through higher fidelity simulations and ITSs, without manually and painstakingly recording the details of a specific performance, often without the use of pre-and post-tests, and without self-reports (Rosen & Foltz, 2014; Hao, Liu, von Davier & Kyllonen, 2015). We are now able to record the actions of multiple test-takers in real time if the assessment takes place in a virtual environment. Moreover, the available technology allows the collection of different types of data, including multimodal data, which can record affective behavior of test-takers during the performance, learning, and assessment process (CPS; see Luna Bazaldua et al., 2015). However, the richness of the data that we can collect is beyond the scope of standard psychometric models. Data from ITSs and performance assessments require different types of analyses, from time-series models and dynamic models from statistics and economics, to algorithms from data mining and machine learning. Recently, von Davier (2015) introduced the concept of computational psychometrics to emphasize the need to blend the data-driven algorithms in a theoretical psychometric framework in order to support accurate and valid measurements.

Process data include the minutia of actions of the test-takers and fine-grain states of the virtual tutoring environment during the interaction of the test-takers with the ITS. The process data are automatically collected and stored into log files in conjunction with the outcomes of the test-takers on the performance tasks. Hao, Smith, Mislevy, von Davier, and Bauer (2016) describe systematic ways to design structured log files.

One challenge in analyzing log file data is determining the meaning of individual actions. There may be some process variables that are relatively easy to measure, such as the identification of a specific path through the different levels in an ITS or the time spent at each task. The number of attempts in solving a problem and the number of hints requested by the test-takers before solving a problem correctly can also be easily counted. However, beyond these kinds of descriptive variables, interpreting actions may be much more complex because of the dynamics of learning and the sheer volume of data generated in log files. In fact, both the cognitive psychologist and the psychometrician struggle with the sheer amount of data and the usual lack of structure found in log files.

The major challenge in analyzing ITS data from a psychometrician's point of view resides in the dependencies and multidimensionality in the data. First, there are dependencies among the responses of the test-takers to the questions or items at each point in time. Some of these dependencies are due to the number of attempts to solve each item, some others are due to the task design, for instance, a reading passage with multiple questions referring to the same passage. Then, there are dependencies across time, which occur due to the fact that the test-taker is expected to learn over time from the interaction with the previous items. These data dependencies invalidate the assumptions of the traditional psychometric models, such as item response theory (IRT) models. Second, the multidimensionality in the data comes from the subskills or knowledge units that comprise and define the domain to be taught in the ITS. Extracting key features from the noise surrounding such data is crucial not only to make analysis computationally tractable (Masip, Minguillon & Mor, 2011), but also to extract relevant features of test-taker performance. One way to attempt to find patterns among these different types of data is to make use of data-mining techniques (Baker, 2015).

The outcome data that are used in ITS for measuring test-takers' skills and subskills are collected through the evaluative scoring. This is the scoring of step-by-step responses of individuals throughout the process. Here we distinguish between the scored-responses and the process of delivering a response. In other words, some parts of the process data are directly scorable and considered outcome data. For example, if the test-takers respond to explicit questions, then those responses are "outcome data". The timing information and

the navigation through the ITS are process data. While we often analyze process data and creative measures that aggregate over observations in ITSs, or examine sequences of observations that match strategies, we may also apply stochastic processes to the process data that were not aggregated, hence the distinction between outcome and process data. Nevertheless, the outcome data exhibit the same dependencies and multidimensionality described previously. For example, a test-taker's actions during the performance task can be scored as correct or incorrect by a human rater or an automatic scoring engine for each attempt. In addition, pretest and posttest data, if available, are individual outcome data. If either of these tests is available, then the test scores that contain information about the test-taker ability can be corroborated with the information contained in the actions scored throughout the task.

Until recently, the methodology used for measuring test-takers' proficiencies and skills in ITSs has been the subject of the field of educational data mining (EDM), computer science, and cognitive science rather than of psychometrics. In the frameworks of EDM and computer science, the most common approach is the Bayesian knowledge tracing (BKT) method (Corbett & Anderson, 1995). In this method, the test-taker knowledge is modeled as a latent variable. The latent variable is updated based on the correctness of the observed responses to the items, which present test-takers' opportunities to apply the skill that is being taught by the ITS (these are the responses to explicit questions, such as multiple-choice items and are called the outcome data in the description above). Test-takers receive attribute-specific feedback based on many practice opportunities over time. The method assumes that knowledge is dichotomized and represented as a set of binary values of variables, one per skill; the skill is either mastered by the test-taker or not after each one particular observation. Observations in BKT are also binary: a test-taker gets a problem either right or wrong, which is used to estimates the levels of proficiency of the test-taker in each of the knowledge units/skills in a Bayesian framework.

BKT is a special case of a stochastic process called the hidden Markov models (HMMs). This HMM instantiation has two latent states, "learned" and "not learned," and describes a learning process. In its common parameterization, it has four parameters: two parameters that describe the learning (the probability that the skill is known before the first item is presented and the probability that the skill is learned after an item is presented) and two parameters that describe the performance (the probability of a correct guess and the probability of a slip). The traditional BKT does not model the potential "forgetting" of a learned material. The HMM can deal with the data dependencies over the course of the ITS. The BKT method relies on the decomposition of the construct to be learned into knowledge units, which is conceptually slightly similar to a learning progression in formative assessments, and it can be as sophisticated as a learning progression.

This step is followed by a mapping of the items and tasks onto these knowledge units, not unlike a Q-matrix in diagnostic models in psychometrics. The test-taker will need to solve correctly a specified number of items from each knowledge unit before moving ahead to the next knowledge unit or next level. The number of attempts and hints that the test-taker receives, the number of responses, or the path of the test-taker through the knowledge units create traces of the test-taker's knowledge. ITSs often uses BKT for mastery learning and problem sequencing and most often BKT has only skill-specific parameters, but Yudelson, Koedinger, and Gordon (2013) extended this model by introducing test-taker-specific parameters. The BKT method assumes that the subskills are independent to address the complications due to multidimensionality. By applying the BKT method to a very large data set, one hopes that violating this assumption will not substantially impact the ongoing estimation of the subskills.

Very few attempts have been made in using any of the traditional psychometric methods on ITS data because of the dependencies among the items and the dependencies over time that are impossible to account for in the traditional models. Some studies considered dynamic Bayesian networks (DBNs) for modeling these data. Other approaches that have been considered are traditional HMMs to model the latent states of learning across different units of knowledge and the probability of transitioning from one state to another.

The Computer Scientist's View

"Building new computational methods and techniques for the development of learning systems and the analysis of data about learning."

From a computer scientist's perspective, much of the early work in ITSs was built on the use of different artificial intelligence techniques. Commonly used techniques coincided with those applied in the development of expert systems such as (fuzzy) production rules in AutoTutor for dialogue-based tutoring of science (Graesser, 2016) and constraint-based modeling in SQL-Tutor for tutoring about the development and use of relational databases (Suraweera & Mitrovic, 2002). The application of these techniques has significantly improved learning outcomes, help seeking behavior, and meta-cognition (Ma, Adesope, Nesbit, 2014). The early intelligent component of these systems came from techniques typically based on symbolic knowledge representation and formal reasoning traditions of artificial intelligence. However, the early ITSs could not handle data about learning experiences accumulated by many learners. They were designed as closed systems whose domain, tutoring, and assessment modules were defined during the development phase and their changes would be difficult to implement.

The field of educational data mining emerged as an attempt to address some of the above limitations of early ITSs. There was a discovery-based analysis of digital traces of learners' activities (also known as log or trace data). Borrowing the foundations from data mining and machine learning, educational data mining has developed methods that are commonly used to address a number of tasks such as prediction performance of learners, identification of strategies commonly used by different test-taker subpopulations, and probability of guessing and slipping rates while studying (Baker & Yacef, 2009). The commonly used methods for addressing various tasks are based on clustering, classification, and association rule mining with the most used methods being decision trees, neural networks, and Bayesian modeling (Romero & Ventura, 2010). The main computational challenge in educational data mining is related to the design of scalable data-mining methods for analysis by incorporating components taken from learning theories and psychometric theories.

Principles established in ITSs have recently been applied in the development of more open learning environments. For example, an authoring toolkit for the development of ITSs can now be used to deploy ITSs onto open-ended learning environments, such as massive open online course (MOOC) platforms (Aleven et al., 2015). As mentioned earlier, the most prevalent method in educational data mining is BKT (Corbett & Anderson, 1995), which now found its applications outside ITSs such as MOOCs. To support such applications, the work of computer scientists on BKT also involved the design of computationally efficient and scalable versions of the BKT algorithm and the construction of practical software libraries that can be used for both the development of learning systems and the analysis of different data sets (Slater et al., in press).

A number of phenomena are still not fully understood about learning with the use of trace data only, although many insights can be obtained from the analysis of log data about the use of ITSs. For example, trace data can be used to detect some of learners' affective states (Bosch et al., 2016) and explain how affective states change as learners are interacting with an ITS (D'Mello & Graesser, 2012). Alternative sources of data about learning with ITSs (and beyond) – such as galvanic skin response, eye gazing, and face recognition – have attracted much attention recently to complete the use of trace data (Calvo & D'Mello, 2010; Azevedo, 2015).

In general, there are two main challenges from the perspective for analysis of such multi-channel data. First, the shortage of standard methods that are used for analysis of different data streams (Sottilare, Graesser, Hu & Holden, 2013). While cognitive psychologists collect huge amounts of data from multiple channels, the challenge for computer scientists is to provide them with methods that go beyond the basic analysis

(e.g., time plots) of data streams from a single channel (e.g., electro-dermal activity). Therefore, feature engineering in EDM for data streams is an important challenge for computer scientists who need to collaborate closely with cognitive psychologists and psychometricians to identify meaningful yet practically useful features from these data streams. Future engineering also needs to make sure that theoretically valid constructs are identified that are of relevance to cognitive psychologists. Second, the need to combine multiple sources of data in the analysis is another challenge of computer scientist in EDM and ITS applications that needs to be addressed. For example, there is a need to triangulate findings obtained from data streams with data collected through content analysis of discourse data collected with either think aloud protocols or group discussions. However, creating a combined and comprehensive analysis model that incorporates different data types is an open research challenge.

Conclusions and Recommendations for Future Research

In this chapter, a cognitive psychologist, a psychometrician, and a computer scientist each presented their view on the most prominent challenge in their respective field when exploiting the large amounts of data collected within ITSs and when considering how this type of data could be used to improve the validity of ITSs. Each of the three researchers did so without knowing specifically what the others were writing and when collating the entire chapter, it came a bit as a surprise that the overlap across perspectives was more pronounced and substantial than initially expected. For instance, in all three perspectives, the large amount of data and the fuzzy relations between the different variables are flagged as a major challenge albeit with somewhat different emphasis depending on the specific view.

This implies that, on the one hand, the discourse is not at its beginning anymore and that substantial exchange between the fields occurs, whereas there is, on the other hand, still ample room for improvement. This is important for the development and implementation of GIFT for several reasons. One of them is that GIFT now has the opportunity to fortify ITSs with more rigorous and advanced assessment tools that are developed under a multidisciplinary perspective. That is, when assessing knowledge, skills, and abilities of test-takers, it is beneficial for the validity of ITSs to incorporate psychometric advances and to implement assessment tools that allow for inferences on the level of psychological constructs and not on specific empirical indicators only including all three perspectives. In addition to this, the development of ITSs within the general GIFT architecture needs to ensure that all stakeholders – and this includes cognitive psychologists, psychometricians, and computer scientists alike – have the chance of making specific and targeted input to achieve the best and most relevant information from several disciplines. To this end, each of the three sections contains a brief summary with outlook and all of these specific outlooks could well serve as an overall outlook of this chapter. So, in concluding, it only remains to say that the kind of discourse exampled in this chapter is already ongoing in the real world and within ITSs, but that – as this discourse intensifies and as borders across disciplines become increasingly weak – the validity of the use of process data as well as ITSs in a more general sense and within a more general framework such as GIFT will continue to increase as we cross borders in interdisciplinary efforts.

References

Aleven, V., Sewall, J., Popescu, O., Xhakaj, F., Chand, D., Baker, R., ... & Gasevic, D. (2015). The beginning of a beautiful friendship? Intelligent tutoring systems and MOOCs. In *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 525–528). Springer International Publishing.

Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist, 50*, 84–94.

Baker, R. S. (2015). *Big data and education* (2nd ed.). New York, NY: Teachers College, Columbia University.

Bosch, N., D'Mello, S. K., Baker, R. S., Ocumpaugh, J., Shute, V. J., Ventura, M., Wang, L. & Zhao, W. (2016). Detecting student emotions in computer-enabled classrooms. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)* (pp. 4125–4129). Menlo Park, CA: AAAI Press.

Baker, R. S. & Yacef, K. (2009). The state of educational data mining in 2009. A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Corbett, A. T., Anderson, J. R. (1995). Knowledge tracing. Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278.

D'Mello, S. K. & Graesser, A. C. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*, 145–157.

Forsyth, C.M., Graesser, A.C., Pavlik, P., Millis, K. & Samei, B. (2014). Discovering theoretically grounded predictors of shallow vs. deep-level learning. In Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B.M. (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014)* (pp. 229–232). International Educational Data Mining Society.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S. J., Toto, E. & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining, 4*, 104–143.

Goldhammer, F., Naumann, J., Stelter, A., Toth, K., Rölke, H. & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*, 608–626.

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education, 26*, 124–132.

Graesser, A.C., Hu, X., Nye, B., Sottilare, R. (2016). Intelligent tutoring systems, serious games, and the Generalized Intelligent Framework for Tutoring (GIFT). In H.F. O'Neil, E.L. Baker, and R.S. Perez. (Eds.), *Using games and simulation for teaching and assessment* (pp. 58–79). Routledge: Abingdon, Oxon, UK.

Haier, R. J. (2016). *The neuroscience of intelligence*. Cambridge, UK: University Press.

Hao, J., Smith, L., Mislevy, R. J., von Davier, A. & Bauer, M. (2016). Taming the logfiles from game/simulation-based assessments. Data models and data analysis tools, Research Report ETS RR–16-10.

Hao, J., Liu, L., von Davier, A. & Kyllonen, P. (2015), Assessing collaborative problem solving with simulation based tasks. In O. Lindwall, P. Häkkinen, T. Koschmann, P. Tchounikine & S. Ludvigsen (Eds.), *Exploring the material conditions of learning. Computer-supported collaborative learning. Proceedings of 11th international conference on computer-supported collaborative learning.* Gothenburg, Sweden.

Luna Bazaldua, D. A., Khan, S., von Davier, A. A., Hao, J., Liu, L. & Zhang, Z. (2015). On convergence of cognitive and noncognitive behavior in collaborative activity. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura & M. Desmarais (Eds.), *Proceedings of the 8th international conference on educational data mining*, Madrid, Spain.

Ma, W., Adesope, S., Nesbit, J. (2014). Intelligent Tutoring Systems and learning outcomes. A meta-analysis. *Journal of Educational Psychology, 106*, 901–918.

Masip, D., Minguillon, J. & Mor, E. (2011). Capturing and analyzing student behavior in a virtual learning environment. In C. Romero, S. Ventura, M. Pechenizkiy & R. S. J. d. Baker (Eds.), *Handbook of educational data mining* (pp. 339–351). Boca Raton, FL: CRC Press.

Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou & L. Jain (Eds.), *Serious games and edutainment applications* (pp. 169-195). London, UK: Springer.

Mitrovic, A., Ohlsson, S. (1999). Evaluation of a constraint-based tutor for a database language. *International Journal of Artificial Intelligence in Education, 10*, 238–256.

OECD. (2014). *PISA 2012 results: Creative problem solving*. Paris: OECD Publishing.

Pardos, Z., Bergner, Y., Seaton, D. & Pritchard, D. (2013). Adapting Bayesian knowledge tracing to a massive open online course in edx. In *Proceedings of the 7th international conference on educational data mining*, Memphis, TN, United States.

Romero, C. & Ventura, S. (2010). Educational data mining. A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*, 601-618.

Rosen, Y. & Foltz, P. W. (2014). Assessing collaborative problem solving through automated technologies. *Research and Practice in Technology Enhanced Learning, 9*, 389–410.

Schneider, W. & Chein, J. M. (2003). Controlled and automatic processing. Behavior, theory, and biological mechanisms. *Cognitive Science, 27*, 525–559.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S. & Gasevic, D. (in press). Tools for educational data mining. A review. *Journal of Educational and Behavioral Statistics.*

Sottilare, R., Graesser, A., Hu, X., Holden, H. (Eds.) (2013). *Design recommendations for Intelligent Tutoring Systems. Learner Modeling (Vol. 1).* Orlando, FL: Army Research Laboratory.

Suraweera, P. & Mitrovic, A. (2002). KERMIT: A constraint-based tutor for database modeling. In *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 377–387). Heidelberg: Springer.

Von Davier, A. A. (2015, July). *Virtual and collaborative assessments. Examples, implications, and challenges for educational measurement*. Invited Talk at the Workshop on Machine Learning for Education, International Conference of Machine Learning, Lille, France.

Yudelson, M. V., Koedinger, K. R. & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. In *International Conference on Artificial Intelligence in Education* (pp. 171–180). Heidelberg: Springer.

SECTION III

GENERAL ASSESSMENT METHODS

Dr. Gregory A. Goodwin, Ed.

CHAPTER 15 – Principles of Assessment in the Generalized Intelligent Framework for Tutoring (GIFT)

Gregory A. Goodwin
US Army Research Laboratory

Introduction

The challenges of assessing human performance have been around for centuries, long before the advent of the Generalized Intelligent Framework for Tutoring (GIFT). This section focuses on a discussion of principles of assessment and on how those principles should be applied in the context of GIFT delivered training. While concepts like reliability and validity are discussed, the goal of this section is not simply to provide a review of basic psychometric terms and methods. Rather, the authors have focused more on applying lessons learned from their respective experiences and disciplines to make recommendations on the development and design of GIFT.

The chapters examine quite a range of topics from the somewhat philosophical (e.g., understanding the reasons and goals of assessment) to the very practical (e.g., developing an instructor dashboard to make student performance measures more accessible to instructors). Key themes that run through these chapters include understanding how evidence and data can be used to automatically support the validation of assessments in GIFT, creating a standardized method for implementing assessments in GIFT, and providing transparency of assessments to users of GIFT.

Chapter Summaries

Why Assess: The Role of Assessment in Learning Science & Society

Benjamin D. Nye, Piotr Mitros, Christian Schunn, Peter Foltz, Dragan Gašević, Irvin R. Katz

This chapter begins with some challenging questions like: Why do we assess? What outcomes and goals are worth assessing? What outcomes and goals are possible and impossible to assess? Are there cases where assessment might be harmful, and why? In answering these questions, the authors explore the trade-space of assessment. All assessments have strengths and weaknesses and understanding those is necessary to choose the most appropriate measure.

Assessments can have unintended consequences. For example, we know that high-stakes assessments that are used to determine pay and promotion for teachers or rankings of schools often pressure instructors to narrowly focus their teaching on the test. Rather than trying to teach students a broad range of skills that they need, instructors drill students on test-taking techniques and on solving the kinds of problems that are likely to be on the test. Fortunately, the authors note, reliance on these high stakes tests are giving away to lower-stakes, formative assessments in educational settings.

In selecting assessments, consideration should be given to four primary factors: reliability, validity, use, and proportionality. Reliability refers to the stability of the measure. Validity and use are related in that they should be aligned. For example, a measure that is designed to address math skills should not be used to determine whether someone should be a mathematician. Finally, proportionality refers to the alignment between the importance of the outcome being measured and the priority of that measure in the instrument.

The authors provide recommendations for assessment design and also discuss the connection between assessment and societal goals, getting at the issue of whether we are assessing things that matter. Ultimately the authors remain optimistic about the role of assessment in intelligent tutoring systems (ITSs) and education more broadly. They see positive trends toward healthier uses of assessment and they believe that ITSs may be part of a trend toward implementing continuous, low stakes, formative assessments that serve a constructive role facilitating learning of critical knowledge and skills.

Assessment of Forgetting

Philip I. Pavlik Jr., Jaclyn K. Maass, and Jong W. Kim

This chapter discusses the importance of several factors including forgetting, task type, and assessment methodology in the measurement of knowledge and skills. When considering task type, the main distinction that needs to be made is the difference between semantic knowledge and procedural knowledge when considering assessment. For example, when assessing semantic knowledge, different techniques have different sensitivities to recall and forgetting. Tests of recognition are usually much easier and less sensitive to forgetting than tests of recall, or transfer.

When assessing procedural skills, techniques will vary depending on whether the evaluation is of a motor or a cognitive procedure. Motor procedures should be evaluated in a way that allows the student to demonstrate the skill with good physical fidelity. Cognitive procedures, like tactical decision making, need to consider the situation in which those procedures are executed. That is, it is important to replicate the conditions that require the skill to be executed, so that the learner also gets practice determining that the cognitive procedure needs to be employed.

Finally the authors talk about the fact that skills and knowledge decay over time and assessments should take this into account. An assessment immediately after training will typically show the maximum recall but over time, performance and/or knowledge declines. The authors point out that, if the goal of the training is to produce long-term recall, GIFT should adjust the training schedule to facilitate long-term retention. For example, spaced practice over a long period of time. On the other hand, if the student is going to practice the skill or use the knowledge in his/her job almost immediately, then the training schedule could be used that does not need to promote long-term retention.

Validity Issues and Concerns for Technology-based Performance Assessments Conclusions and Recommendations for Future Research

Irvin R. Katz, Michelle M. LaMar, Randall Spain, Juan Diego Zapata-Rivera, Jo-Anne Baird, and Samuel Greiff

This chapter provides an excellent discussion of the complex nature of validity. One of the primary themes of the chapter is that there are many ways to measure validity and the way that one might choose depends on the claims being made about the assessment. As the authors put it: “Assessment is a form of evidentiary argument, for which evidence is collected in support of a particular claim that is defined by the testing purpose”.

To illustrate this point, consider the Scholastic Aptitude Test (SAT). This test, taken by millions of high schoolers hoping to go on to college offers scores on dimensions like math and verbal ability. If the claim being made about these scores is that the scores reflect an overall math or verbal ability, then one would validate that claim by looking at student performance on a representative set of tasks that depend upon those respective abilities. On the other hand, if the makers of the SAT claim that it predicts grades in the first year

of college, then that claim would be validated very differently (for example, by looking at the correlation between SAT scores and freshman year GPAs.

Technology provides unique ways of conducting assessments that were inconceivable until recently. Technologies such as computer adaptive testing, natural language assessments, game-based assessments, and more, have both opened exciting new opportunities for assessment and new challenges for the validation of those assessments. In the last section of this chapter, the authors discuss some of the challenges for validation that are created by these technology based assessment methods.

Toward Systematic Assessment of Human Performance Interventions in the US Army: An Assessment Process Framework

Kara L. Orvis, Jared M. Freeman, Jeffrey M. Beaubien, Clayton W. Burford, Joan Johnston, Lauren Reinerman-Jones, Grace Teo

This chapter describes an assessment process framework (APF) to standardize the conduct of training assessments within the Army. The APF was developed to be broadly applied across the evaluation of technologies, studies of training interventions, the introduction of new work procedures and policies, the adoption of new organizational structures, and so forth. At the end of the chapter, the authors outline how it could be useful for planning assessments in GIFT.

The APF identifies the major steps for planning and executing an assessment. This process begins with the receipt of a task to conduct an assessment. The assessment team then plans the assessment by first refining the goals and framework. This entails identifying the hypotheses to be tested and/or the question to be answered. It also includes the identification of a theoretical model to predict the effects to be observed. Next, the team defines its measurement methods. Once the plan is complete, the team should develop materials and pilot test the measures to insure that everything works as expected. The final design and materials would then be approved by the sponsor and then the assessment can be conducted.

In the final section of the chapter, the authors describe how the APF could be applied to developing assessments in GIFT. Currently, there is no framework for a course developer building assessments in GIFT. It is just assumed that the course developer will know what to assess and how, but this may not be a valid assumption. Providing a framework and perhaps integrating that into the GIFT authoring tool has the potential to insure better quality control for the development of assessments in GIFT.

Assessment in Intelligent Tutoring Systems in Traditional, Mixed Mode, and Online Courses

Anne M. Sinatra, Scott Ososky, and Robert Sottilare

This chapter discusses possible ways of using GIFT for instruction in traditional, mixed mode, and online courses. For example, GIFT might be used to provide a block of instruction in a classroom or it might be used for remedial training for students that are having trouble with concepts in the class. It might also be used to provide an in-class quiz or to provide preparatory instruction for an in-class lesson. There are many other possible uses of GIFT in these different modes of instruction but all of them fundamentally boil down to being able to use the assessment capabilities of GIFT.

Currently however, GIFT does not have a convenient, easy to use means of viewing the results of learner performance and interactions. This highlights the need for an instructor dashboard. Though such a dashboard does not currently exist in GIFT, the authors describe a notional user interface and functionality that

could be associated with an instructor dashboard. Some of the measurements that an instructor dashboard would facilitate include measures of learner progress through the material, learner attitudes about the training, learner behaviors and performance, and interactions among learners when doing group exercises to name a few.

Lessons Learned from Large-scale e-assessments: Future Directions for GIFT

Jo-Anne Baird, Anne M. Sinatra, Gregory Goodwin

Computerized assessment was embraced in high-stakes testing as soon as it was available. Originally, punch cards were used for statistical processing, optical scanners for processing multiple-choice answer sheets, and mainframe computers for handling the huge amounts of data processing and reporting. Unfortunately, despite decades of development of computer technologies, the use of computers in large-scale testing has evolved very little. Despite the possibilities of using complex virtual environments, games, and simulation to evaluate skills and knowledge in completely novel ways, computers are still largely used to deliver the same sorts of assessments that were traditionally given on paper and pencil tests.

Two of the biggest challenges associated with current large-scale computerized tests are transparency and generalizability. Transparency refers to the ability of the test-takers to understand the basis of the evaluation. Traditionally, large-scale test developers fought to keep their questions secret so that they could be reused and still maintain their validity. Test-takers were not shown the questions they answered correctly or incorrectly so they had no way to challenge the assessment of the exam. Just as in the civilian world, in the military, any system that provides an evaluation without offering a basis for that evaluation is unlikely to be trusted or used. Especially if the evaluations are suspect or have a negative impact on the careers of service members.

Generalizability, the other challenge associated with large-scale testing, has to do with the transferability of the measure. That is, do the assessments generalize to real-world performance? As noted in other chapters, this question is similar to the question of validity. Specifically, what evidence is there to support the claim that the assessment predicts/reflects performance in real world contexts.

The use of GIFT in live or simulation based training events has both great promise and potential risks. It is possible that GIFT could use very sophisticated algorithms to provide performance assessments; however, if those assessments lack transparency, leaders will be unlikely to pay attention to them. Somehow GIFT will need to be able to provide evidence of the generalizability, transfer, and or validity of the measures.

Recommendations for GIFT

A theme that comes through all chapters in this section is the importance of the validation of measures in GIFT. Validation is described as an evidence-based conclusion. That it, the claims of an assessment, whether it is to make a prediction or measure a construct must be supported with some kind of evidence. Not all validations are created equally. For example validating a measure of declarative knowledge (e.g., does the learner know the state capitol of Florida?) can be verified with a simple multiple-choice question. On the other hand, validating a measure of leadership ability may be much more challenging. Currently, most assessments in GIFT are at a fairly low, declarative knowledge level but over time they will need to measure increasingly complex competencies in learners.

All of the chapters envision tools that GIFT can use to facilitate and automate the process of validating these assessments. As GIFT evolves, it would be good to provide validation tools to content developers. For example, if GIFT were to have access to performance measures that were available from databases at

training sites, it might be possible to automatically examine the validity (or transfer) of skills learned in GIFT to other, more advanced live training or operational environments. Some of this validation data should probably be presented in the instructor dashboard.

Additionally, GIFT should be able to generate measures of question reliability by cross-referencing related questions. For example if some items are poorly correlated with other items that assess understanding of a concept, then the inconsistent measure should be flagged as unreliable so that it can be removed from the test bank.

Another means of improving the quality of questions is the incorporation of the APF, presented by Orvis et al., provides a standard method for developing assessments that could be used in GIFT. This is a tool that would be useful for course developers and the incorporation of this tool into the authoring interface might help to insure a consistent quality for assessments developed for GIFT. Future research should consider ways in which the assessment process framework, or at least some of its components, could be incorporated into the GIFT interface.

Ultimately, these tools will need to provide a means of addressing the three questions of validity, generalizability, and transparency of GIFT's assessments. That is, they will need to enable students and instructors to understand the basis or evidence for the assessments made by GIFT. Adding these capabilities to GIFT will be instrumental in insuring a widespread adoption of this system.

CHAPTER 16 – Why Assess? The Role of Assessment in Learning Science and Society

Benjamin D. Nye[1], Piotr Mitros[2], Christian Schunn[3], Peter W. Foltz[4,5], Dragan Gašević[6], and Irvin R. Katz[7]
USC, Institute for Creative Technologies[1], EdX[2], University of Pittsburgh[3], University of Colorado -Boulder[4], Pearson[5], University of Edinburgh[6], Educational Testing Service[7]

Introduction

Why do we assess? What outcomes and goals are worth assessing? What outcomes and goals are possible and impossible to assess? How do we use assessment to improve all outcomes? Why has assessment become controversial in recent years? Are there cases where assessment might be harmful, and why? What are sources and levels of error in assessment and how do those impact individual students, schools, and society? Are there times when we might not wish to assess certain outcomes?

Even though assessment often is imperfect, it provides valuable input to the process of teaching, learning, and educational resource design. However, narrow assessment, especially used in high-stakes settings, can lead to worse educational outcomes (e.g., performance in later courses, workplace, or social settings; Hout & Elliott, 2011). Teachers may have a strong incentive to teach to the test, leading to a strong focus on memorization and rote procedural knowledge, while compromising key skills such as empathy, groupwork, mathematical maturity, and analytical reasoning. These are thorny problems – education shapes the skills[1] that shape society, so these questions have broad implications. With that said, by constraining the discussion to the kinds of constructs considered when building learning experiences, the goals of assessment become more tractable.

To fully consider the role of assessment for learning technologies and intelligent tutoring systems (ITSs), we must first consider the role of assessment in general. Fundamentally, educational assessment measures relationships that influence the learning process and its outcomes (Gipps, 1994). These measurements are intended to help make better pedagogical decisions to achieve learning and behavioral outcomes. Traditionally, the results of these assessments are leveraged by students, teachers, administrators, policymakers, employers, and other educational stakeholders. More recently, artificially intelligent machines such as ITSs use assessments to achieve their goals. The types of assessments and skills included in ITSs have traditionally been fairly uniform (e.g., math problem solving, recall) when compared to the broad range of potential assessments (e.g., peer review, team performance, complex simulations unfolding over hours or longer). With that said, ITSs have recently been growing to accommodate a greater range of assessments, such as interactive dialog-based assessments, assessments of physical tasks such as marksmanship, and assessments of project teams for engineering classes (Nye, Goldberg & Hu, 2015; Rosen, Ferrara & Mosharref, 2016; Chesler, Ruis, Collier, Swiecki, Arastoopour & Shaffer, 2015). This raises the question of what the next generation of ITS architectures should assess.

This chapter considers the multiple roles for assessment of and for learning, the (sometimes competing) high-level outcomes we seek to measure, and how local measures of learning connect to broader societal goals. We begin with a discussion of "what gets measured gets done", how the consequences of assessment

[1] While this chapter frames assessment in terms of measuring skills, we recognize that educational assessment covers a wide range of constructs, from invariant traits to attitudes toward different academic domains. We believe that the majority of this discussion also applies to this broader conception of assessment, but focus on skills/knowledge as this is the primary focus of educational assessment.

plays out in the classroom (sometimes leading to an emphasis on superficial skills rather than deeper instruction), and approaches to mitigate this. These issues lead to a consideration of how we decide what to assess, and the role played by newly recognized 21st century skills, such as persistence, digital literacy, and teamwork. Then, connections between traditional skills, 21st century skills, and longer-term emergent outcomes are discussed. We conclude with recommendations for generalized ITSs in terms of the domains they assess and how skills are measured across different time horizons, such as formative during learning, summative after learning, future on-the-job performance, social group/team performance, as well as with a discussion about strategies surrounding important outcomes that cannot yet be measured.

What Gets Measured Gets Done: Intended and Unintended Consequences

The primary benefit of educational assessment, like many assessments, is the old adage "what gets measured gets done," attributed to Lord Kelvin (1883). Assessments can serve a variety of positive functions in education, relevant to either individual learners and/or for aggregates (e.g., classes, districts):

1) Identifying areas of competence or weakness for specific topics (e.g., certification).

2) Evaluating instructional strategies that could be replicated in other contexts.

3) Tracking improvement over time.

4) Adapting to student or class behaviors in real time or between lessons.

As such, assessments play a critical role for monitoring and improving the learning process. When compared to real-life tasks, assessments tend to cover a lower-dimensional set of skills (i.e., simplified tasks) or a lower-dimensional range of contexts (i.e., assessed under a reduced range of conditions). Despite this, even simple assessments may measure skills that are required for success in an extremely broad range of real-life tasks. For example, the ability to pass a simple literacy test implies basic fundamental skills needed for a vast number of careers and other roles in society. Traditionally, assessments for education tend to focus on preparation for a broad range of later experiences, which implies an emphasis on generalizable skills or competencies. On the converse, assessments for training (e.g., for an on-the-job task or piece of equipment) may instead rely on assessments that emulate specific tasks and their environment as closely as possible. In both cases, a variety of assessment tasks are possible. Some common examples of assessment tasks[2] are shown in Figure 1.

However, even when carefully designed, assessments are often weak reflections of the actual skills and knowledge that we want students to master. This is necessary for many traditional assessments designed to identify knowledge gaps, where measurements should ideally identify individual skills and many observations of applying the same skill are desired. Likewise, from the standpoint of instruction, assessments that simplify a problem can help to scaffold a limited subset of skills by practicing on highly simplified tasks (e.g., analogous to how athletes practice simple drills before coordinating them into a competitive game). In both cases, assessments from the lower-left of Figure 1 are logical: a large number of skills may need to

[2] There is a great degree of overlap between these different kinds of assessments, and different types can also be combined (e.g., adaptive simulation-based assessments, open response situational judgement tasks, and peer-assessed portfolios). In general though, these assessments when used in practice tend to not be less complex than shown in Figure 1 (e.g., peer assessment is seldom used for simple recall selected-responses, open responses such as essays or open problem solving tend to require integrating more skills than multiple-choice questions).

be measured quickly and reliably, with minimal confounds due to context. Unfortunately, when used indiscriminately and exclusively, such simplified assessments can lead to optimizing for lower-dimensional models for skills that do not align to real-world tasks or outcomes.

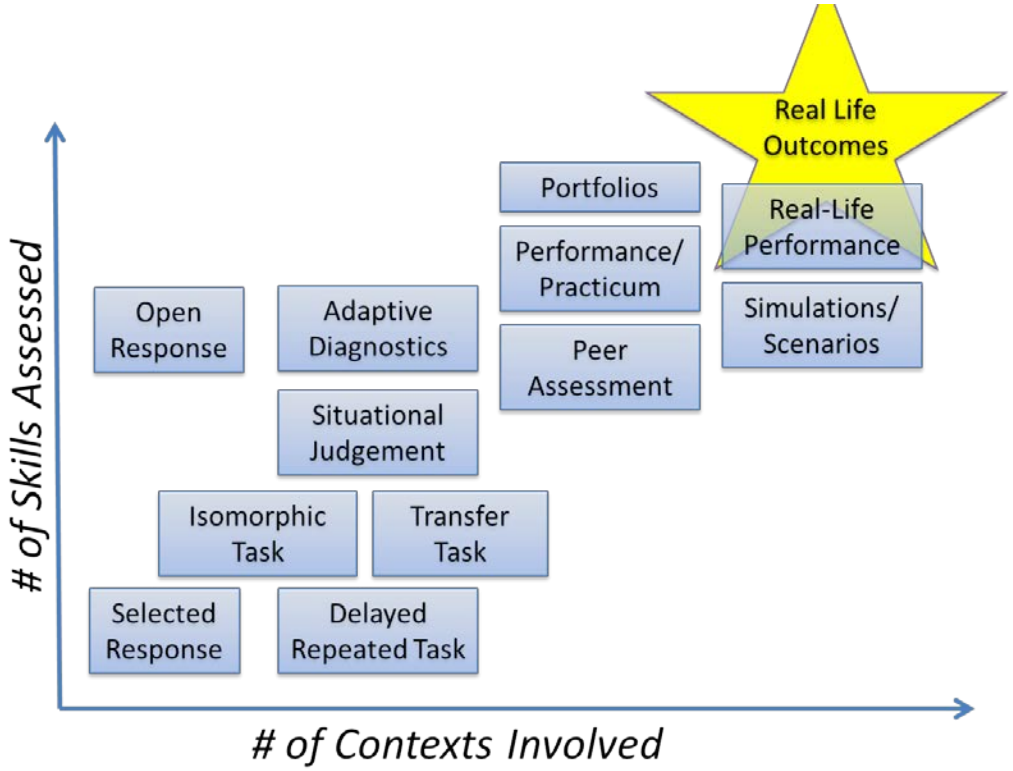


Figure 1. Number of skills and context involved in different assessment types.

In Figure 1, the axes roughly indicate how assessment types are often used in practice, in terms of the number of skills they tend to measure simultaneously and the breadth of contextual information integrated into such assessments. Brief assessment tasks (e.g., multiple choice) often use abstract problems with little context, for maximum reusability and reliability. As assessments use more contextual information, they may leverage different but related tasks, the same task at different times/places, or different situations containing a mixture of both relevant and irrelevant/incidental information to a task. Contextualized tasks might also involve a series of related or interconnected subtasks (common in real-life applications, performances, or simulations).[3]

Stereotypical educational assessments are standardized multiple-choice exams taken annually (or even less frequently) and which are used to make inferences about student, teacher, school, and school district performance. Such exams are very time-limited relative to the breadth of topics and skills they must assess. This means that they can either assess complex skills, such as creative problem solving or give statistical significance through repeated measures, but not both. A similar tradeoff can be found in assessing skills in context-free problems versus assessing the same skills in the context of rich, authentic problems (Anderson & Schunn, 2000). Further, when stakes are high, policies tend to be more conservative about leveraging

[3] In this discussion, "varied contexts" means the conditions under which assessments test for certain skills. It does not have any implications about the range of contexts where the skill may be applied, which can be very broad, even for very simple assessments (e.g., a multiple-choice test on basic addition).

innovations in test design, particularly when it impacts evaluations at the teacher, school, or district levels. For example, while adaptive testing has the potential to help optimize traditional tests of student knowledge (higher precision in fewer items), this is still not ubiquitous more than twenty years after initial explorations of its practical feasibility at scale (Mills & Stocking, 1996). As a result, while such tests are useful for identifying areas of competence/weakness and, in some cases, effective pedagogical strategies (e.g., value-added classrooms), they occur too infrequently and are usually too shallow to be useful for tracking individual improvement or adapting instruction.

At the opposite end of the spectrum, we have seen an increased shift toward low-stakes continuous formative assessment. Continuous assessments allow for a more-varied set of assessment types (more time to apply skills) and also repeated-measures from similar tasks over time or contexts. These have been designed to help students to gain a better understanding of what they do and do not understand (metacognition) and also for instructors to calibrate their pedagogical content knowledge (Black, Harrison, Lee, Marshall & Wiliam, 2004). With growth of educational technology, systems for assessment have become increasingly integrated into classroom experiences, and increasingly rapid in providing feedback (Schell, Lukoff & Mazur, 2013). Systems have also been moving toward gamification and open student models – visualizing and rewarding student progress individually or relatively to peers to motivate students to persistently engage with the system (e.g., Brusilovsky, Somyürek, Guerra, Hosseini & Zadorozhny, 2015).

ITSs can contribute to continuous assessment and increase learning gains both through user-adaptation (VanLehn, 2011) and also through reports to instructors and students produce learning gains (Black et al., 2004). As a side effect, such systems also tend to allow for mastery learning and self-paced learning, which can be beneficial for both learning and student affect (Kulik, Kulik & Bangert-Drowns, 1990). Since they do not punish students for mistakes in the same way as traditional human-graded homework assignments do, they may also reward intellectual risk taking, which may avoid the traditional association of schoolwork with punishment, and allow for more intellectual diversity. This is particularly important because research on adaptive systems indicates that optimal learning may occur when students are less risk-averse, allowing them to encounter desirable difficulties (Tang, Gogel, McBride & Pardos, 2015) and space their practice of different skills over time (Pavlik, Bolster, Wu, Koedinger & Macwhinney, 2008).

That said, even when effective assessments are used, we have seen pitfalls at all levels with reward structures based on many assessments. When school funding and teacher performance is tied to test outcomes, teachers have a strong incentive to teach to the test and may abandon more-effective pedagogy and assessments (e.g., realistic scenarios) in favor of lessons that exclusively align with high-stakes test items (Hout & Elliott, 2011). In terms of Figure 1, this essentially amounts to aligning in reverse: rather than aligning classroom teaching and assessment to the complex real-world skills we want, teachers are incentivized to align backward to reliable but not very useful tasks that capture a limited subset of such skills. Similarly, when students are provided continuous formative feedback and continuous assessment, a subset of students tend to game the system (Baker, Walonoski, Heffernan, Roll, Corbett & Koedinger, 2008). They will often exhibit satisficing behavior, and do the exact minimum required to pass, advance, or achieve the desired outcome, although the design of the assessment tasks can have a large effect on whether such gaming behavior occurs (Kessler, Stein & Schunn, 2015). In high-stakes assessments, such as college entrance exams, this can be seen through high levels of not just gaming behavior, but explicit cheating (Mehra, 2012). Despite these issues, strategies exist that can mitigate some of the risks of assessment, which is discussed in more detail in the following section.

Selecting Assessments: Methodology to Determine How and What to Assess

Theory on how to develop assessments is very broad, and a comprehensive overview is beyond the scope of this work. With that said, we here briefly summarize four factors which determine how assessments are

developed and considerations about what should be assessed. When developing assessments, the two primary considerations for tasks have traditionally been *reliability* (repeatable) and *validity* (measures what it is intended to measure). Modern assessment practices also consider the *use* of assessment (Darling-Hammond et al., 2013). Finally, when considering use, it becomes vital to consider how multiple assessed constructs are weighted *proportionally* to the importance of the skills to some ideal real-world outcomes. To summarize, a well-designed assessment regime should consider the following:

- **Reliability:** Whether an assessment is repeatable and will give similar results when repeated across time and across different test-takers.
- **Validity**: Whether an assessment is measuring a meaningfully interpretable construct.
- **Use:** What purposes the assessment will be used for in practice.
- **Proportionality:** Whether the assessment measures and balances each construct relative to their level of importance to the goals for their intended use.

Reliability

Reliability is primarily a question of measurement noise and statistical significance: less-reliable measures will require more samples to derive the same confidence for inferences. Reliability also considers issues of eliminating items that give bad evidence and of bias. For example, if better students answer a question incorrectly, that is a good indication there is an error in the assessment. Similarly, reliability measures can be used to analyze relative levels of bias or discriminability by learner subgroups, such as reading ability, age, gender, and ethnicity. Certain domains, such as math and reading, have a long history of developing reliable test items, such as those used on the GRE, SAT, ACT, and similar tests. This is partly due to the facts that a large amount of data can be collected and constructs are very carefully defined, allowing consistent measurement. In addition, assessments have incorporated open-ended responses such as writing where the rubrics can still be constrained to ensure reliability (e.g., Foltz, 2016).

Other domains, such as creative writing, are not easily tested using reliable metrics: assessments about the quality of a short story writer might require a portfolio of stories, take many readers, and change over time. Instruments designed for seemingly-fuzzy constructs such as creativity do exist, such as the Torrance Tests of Creative Thinking (Kim, 2006) and recent work toward assessing general problem-solving skills (Greiff, Wüstenberg, Csapó, Demetriou, Hautamäki, Graesser & Martin, 2014). For some other skills such as critical thinking, text mining of student discourse can be used for assessment (Kovanović et al., 2016).

A significant gap for many under-assessed skills is the challenge of developing simple, fast assessments that are suitable for automated grading (a requisite for most standardized testing and also a critical time-saver in a classroom context). That said, continuous assessment of complex tasks, performance-based assessment, and peer assessment are all emerging as potential options for assessing skills that have traditionally lacked reliable assessments (Mitros, 2014). In many cases, these advances are possible either by leveraging more data (e.g., fine-grained analysis of a tutoring scenario; Segedy, Kinnebrew, Goldberg, Sottilare & Biswas, 2015) or by using structured methodologies to derive reliable assessments from either multiple crowdsourced items on a construct (e.g., Mitros, 2015) and/or multiple peer assessments on a learner's response to an item (e.g., Luo, Robinson & Park, 2014). That said, while reliable assessments may be challenging to develop for some skills or outcomes, the fundamental science for reliability is relatively mature.

One hurdle for reliability, gaming the system, is particularly relevant to ITSs, which often provide hints that can sometimes be abused (Baker et al., 2008). One recommendation for avoiding this behavior is to

diversify assessments to include both human and machine assessment, leveraging social dynamics to help control gaming behavior. Likewise, carefully controlling stakes associated with assessment may also reduce incentives to game the system. For example, at one extreme, assessment can be purely formative, even allowing students to override measurements (e.g., mark things as known/unknown). This is common in systems where assessments exist to help students self-regulate. Where such alternate assessments are appropriate, these may help increase reliability even in the presence of learners who would otherwise use the system improperly.

Validity and Use

The validity of an assessment is a more complex issue, since it involves both an objective and a subjective component. Kane (2013) frames validity in terms of the strength of evidence that an assessment gives for one or more inferences. These inferences might be, for example, about an expected theoretical construct (interpretability), about some future event (predictive value), or about their value to influence some decision or action (usefulness; Darling-Hammond et al., 2013). A variety of methodologies exist to gather evidence for validity, some of which rely on theoretical assumptions and others on accumulated empirical evidence. In all cases, validity lies on a continuous scale of the level of confidence that a given assessment gives for an inference. This is further tempered by the expected use of the assessment, in that certain uses may require very high (or very low) levels of confidence and that an assessment may be more valid for some uses than for others.

Aligning to theoretical constructs (interpretability) traditionally relies on coherence and agreement, where multiple raters determine that the results of an assessment fit a certain shared construct. It may also involve agreeing with prior assessments intended to measure the same construct. An initial step in exploring assessment validity can involve cognitive task analyses with domain experts and also cognitive interviews, in which pilot participants restate the assessment items in other words and explain their choice rationales (Leighton, 2004). There are also related strategies for building items with higher validity, such as asking open-ended questions on pilot assessments prior to building final assessments that use multiple-choice questions (Thissen-Roe, Hunt & Minstrell, 2004).

That said, from a practical standpoint, all assessments should ultimately align to real-world outcomes: either they should directly measure real outcomes (e.g., performance or behavior on some space of relevant tasks) or should provide useful inferences about such outcomes (e.g., interpretation or indirect prediction, often based on theoretical or substantive arguments). The first approach, championed by Wiggins (1990), is to use authentic assessments: ones that align closely to real-life conditions. This has a strong appeal in that it ensures that skills should be valid in at least some space of realistic contexts. However, an authentic assessment is not always practical or desirable. First, authentic assessments may be much more time consuming to conduct, resulting in fewer assessment observations and less evidence. Second, authentic assessments may introduce confounding information required to anchor the problem in a realistic instance, but can introduce cultural confounds such as certain learners being unfamiliar with the context used to anchor the task (e.g., a math word problem on baking french bread is unlikely to translate to rural Afghanistan; Nye; 2014). Third, in many contexts, one wants to measure or practice an isolated component of a real-world skill. This is due to the fact that the authenticity of an assessment is partly subjective, depending on the frame of reference for defining a “real-life” task (Gulikers, Bastiaens & Kirschner, 2004).

The second approach to alignment is based on inferences between the assessment to real-life outcomes and performance (Darling-Hammond et al., 2013), even if such inferences require a chain of inferences through other assessments (e.g., prerequisites, analogues, preparation for future learning). Such chains may be theoretical, but are ideally empirical (i.e., derived from data). Two common ways to collect validity data about how an assessment predicts behavior are longitudinal analysis and the ability to discriminate between

groups of experts and novices. In the longitudinal case, earlier assessments may be studied for their ability to predict later real life outcomes, such as success at on-the-job tasks. In the group case, assessments tie their inferences to the assumption that performing similarly to an expert on an assessment implies a greater capability to perform tasks and behaviors like those experts. Both reference points for inferences have trade-offs: longitudinal data can be slow to collect so assessments or even skills may become dated, while group-based data does not necessarily guarantee that the assessed skills are necessary or sufficient to act as an expert.

In practice, assessment validity is estimated using either a psychometric approach or a data-mining approach. In the psychometric approach, assessments are carefully designed for an optimal measurement of one or more constructs, paying careful attention to validity and reliability under models such as item response theory (IRT; Embretson & Reise, 2013). In an educational data-mining approach, the learning experience is not always designed for assessment, but assessments can be mined out of traces of learner activity (Shute & Kim, 2014). ITSs often use hybrid approaches of the two, which can assess behavior at the step level (VanLehn, 2011).

In terms of validity, stealth and purely data-mined assessments can have downsides. First, teachers may not wish to use the assessments because they have low face validity (i.e., they do not seem to be measures of what they claim to measure). Second, data-mining can produce assessment indicators that are very specific to the particular tool and context that might not generalize to other contexts. Third, data-mining often requires larger dataset to validate the items, to reduce the risk of finding "fluke" predictors from a potentially infinite sea of predictors. One workaround for these issues is to align such bottom-up assessments to existing top-down measures to establish consistency (e.g., a separate test used for validation purposes only).

Another significant shortcoming for assessment validity (across all types of assessments) is often the lack of useful measurement of real-life outcomes such as job performance, life satisfaction, societal outcomes, and preparation for future learning. Lacking such data can lead to substantial guesswork, such as entire required courses or assessments that may show great validity in assessing constructs that are not useful because they are unactionable (e.g., correlated but not causal to later outcomes) or even entirely unrelated to later outcomes. Integrating ITSs into on-the-job tasks or realistic simulations with high predictive validity for such performance could provide a wealth of data to back-propagate and inform the validity of simpler assessments. Likewise, artificially intelligent assistants connected to ITSs may someday lend insight into longer-term life outcomes that are associated with certain assessment performance (e.g., identifying the effects of financial literacy tutoring through Amazon purchases).

Proportionality

While not always noted in assessment literature, the concept of proportionality for an assessment program is also essential. In part, this is because proportionality does not necessarily apply to a single assessment item instead applies a group of related assessment items meant to support multiple inferences (e.g., skill levels). In short, since some skills and outcomes are more important than others, it is sensible to prioritize assessment in reasonable proportion to the importance of each desired outcome. For example, when teaching electricians, it would not be sensible to base certification equally on hands-on expertise, knowledge of electrical circuits, and particle physics, since theoretical physics plays a relatively small role in that competency.

While this seems intuitive and obvious initially, proportionality breaks down quickly and silently in practice. First, due to reliability or validity issues, many critical skills and outcomes may have no fast or reliable assessments. Second, the relative importance of skills is often subjective or simply unknown (e.g., fre-

quency of use or consequences of failure never measured rigorously). Finally, due to issues such as prerequisites and co-requisites, certain skills that are not intrinsically important may be pivotal stepping stones to more advanced skills or outcomes. While the last issue can be addressed quantitatively, assuming enough data, the first two are socio-technical problems. Unfortunately, when these challenges exist, too often the solution is to ignore outcomes that lack reliable, valid assessments and instead *invalidly base all decisions on only the assessments available*. Since many essential skills lack fast and reliable assessments (e.g., complex problem solving, creativity, curiosity), this is analogous to having poor visibility out of your car windshield so you just watch your speedometer instead.

Of all issues in assessment, this can be the most subtle, since it is an error of omission. If assessment design is not done proportional to importance, it is easy to leave out key assessments. If assessment design is done proportional to importance (i.e., outcome importance established prior to selecting assessments), then it should be obvious that key skills and outcomes lack measurements. In many cases, this might not be a solvable problem: reliable assessments may not exist, may not be cost effective, or may not be reasonable within time constraints. However, if any of these are the case, then it means that the confidence of the overall set of assessments for a given use must be downgraded as a result (e.g., in the car example, one might drive very slowly in reverse). Without this practice, it can be very easy to over-optimize for a set of relatively unimportant assessments at the expense of more important ones that are not currently measured (but might have been in the past, even if subjectively).

Recommendations for Assessment Design

When selecting assessments from this broader set, we would suggest the following guidelines:

1) For simple procedural knowledge and rote memorization, traditional assessments may be sufficient (bottom-left of Figure 1). Any deeper set of skills should ideally be measured using multiple types of assessment to enable both realistic application and simplified diagnostics.

2) a) Where direct assessment is impractical, data-mining techniques such as stealth assessment may be applied or b) indirect assessment can also be done using a combination of qualitative instructor assessment and/or peer assessment, paired with quantitative assessment of classrooms and schools.

3) If assessments cannot be designed to evaluate distinct skills, alignment may be established at the task level (e.g., simulating real-life tasks) or at the process level. For example, one can look at the alignment of the curriculum design and classroom instruction to evidence-based best practice. Classrooms may also be evaluated be peer assessment by other instructors.

4) Assessments must have a clear (diagramed theory or statistically supported) chain of connections that demonstrates how that assessment provides *useful inferences* for a real-life outcome.

5) Sets of assessments should be proportional, meaning that they also show this chain, but with the added requirement of the level of importance of each of multiple outcomes for the assessment goals. Outcomes that cannot be assessed properly must not be removed, and their uncertainty should be considered rather than ignored.

This line of thought leads to our final central questions: What is worth assessing? Why are we assessing? This process starts with desired learning outcomes, with recent candidates including a particular focus on complex 21st century skills, such as persistence, digital literacy, and teamwork. Once the set of learning outcomes is defined, it is possible to find assessment techniques for some of those outcomes, but not for others. Toward this end, the next section briefly considers the role of education and assessment with respect to broader societal outcomes.

Emergent Outcomes: Connecting Individual Assessments to Societal Goals

As a background for this chapter, a session on this topic with a cross-section of approximately two dozen education experts in different fields and different roles (e.g., researchers, software developers, teachers) identified a set of over 75 distinct learning outcomes, from a broad set of perspectives (Nye, 2016). Two questions framed an intense discussion: "What is worth assessing?" and "How do readily-measurable assessments connect to emergent and societal outcomes that we care about?" The focus was not on specific subjects, but on capturing qualitatively different behaviors and outcomes to track.

Outcomes were first brainstormed, then sorted into four categories based on their grain size for assessment: Near-Term, Emergent/Intermediate, Societal, and Big Picture. Near-Term outcomes included brief assessments (e.g., tasks) and also measurement methodologies that support assessment. The Emergent/Intermediate category covered outcomes that can only be assessed by monitoring patterns across time or contexts. Societal outcomes included results or learner characteristics perceived by participants as beneficial at a cultural level (subject to the lens of cultural values of the participants). Finally, Big Picture outcomes represented high-level ideals stated by participants. The three key Big Picture outcomes identified during the session were summarized as *Adapting to New Things*, *Communication*, and *Happiness/Utility*, which appeared to have some consensus as being intrinsically valuable (as opposed to instrumentally valuable to reach some other outcome).

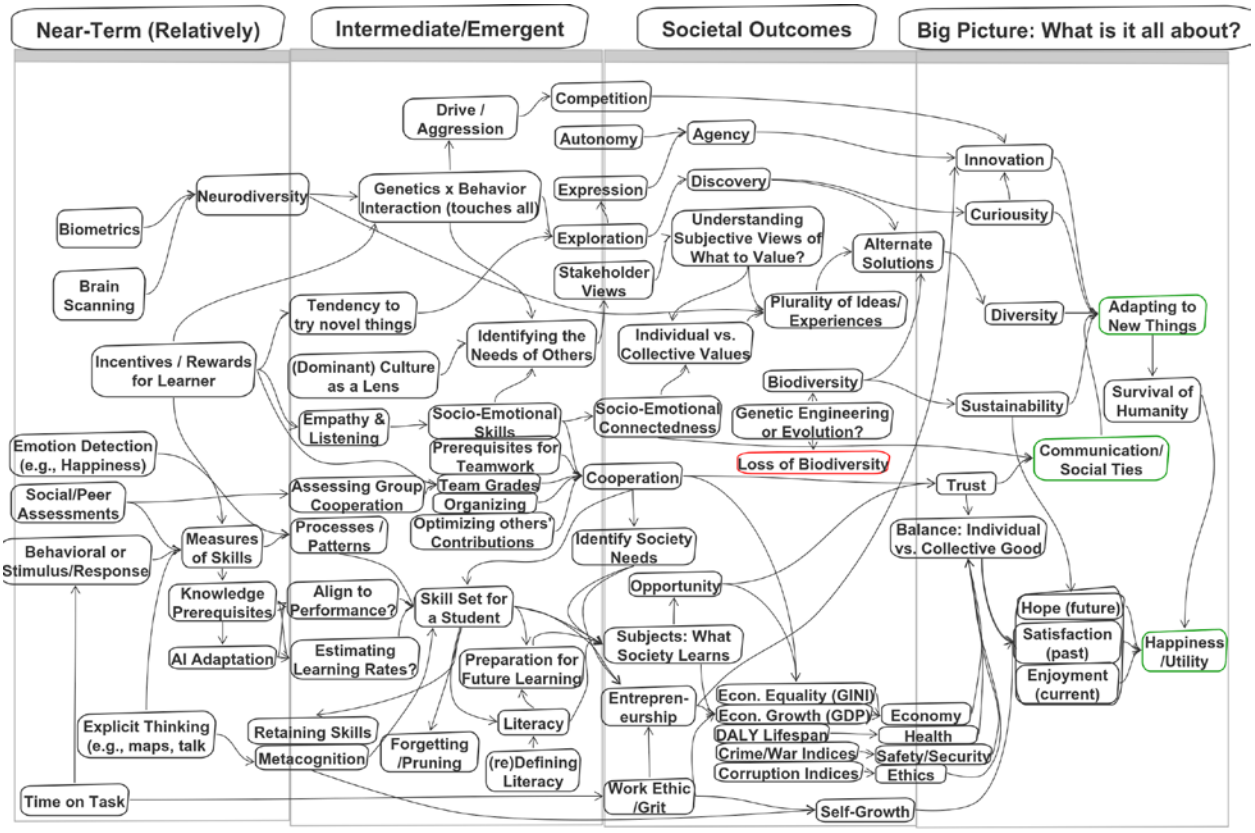


Figure 2. Influence map derived from an expert brainstorming session.

In Figure 2, a set of desired outcomes and assessment methods of those outcomes, as identified by a diverse group of experts. The outcomes are sorted on a scale from immediately measurable, up to highly longitudinal and emergent. This map gives an idea of why assessment is a complex and sometimes controversial

topic – while it is possible to objectively discuss individual outcomes, the prioritization of such is inherently subjective, and different cultures and communities place different importance on each of these.

In the rough influence map inferred from the results of this session shown in Figure 2, a number of pathways in this graph were the result of notable discussions:

1) *Diversity*: The upper-left measurements (*Biometrics* and *Brain Activity*) connected with *Neurodiversity*, nature/nurture interactions (*Genetics x Behavior*), and ultimately, a broader discussion on the dual nature of *Diversity* at both the biological level (*Biodiversity*) and of ideas (*Plurality of Ideas*). In both cases, the ultimate goal was *Adapting to New Things* (not necessarily just by people, but also by ecosystems). There was some consensus that diversity had an optimum where specialization and uniqueness was balanced against the need for communication (i.e., avoiding a Tower of Babel collapse), but with no known metrics to measure this balance.

2) *Curiosity*: The concept of *Curiosity* was thought of as measurable through *Exploration* (looking for new things) and *Discovery* (finding new things), which were both considered a function of overall novelty-seeking (*Tendency to Try New Things*). These skills were thought of as measurable, but not measured commonly enough.

3) *Innovation*: *Agency, Competition, and Entrepreneurship* were all considered central to this outcome. *Agency* and *Autonomy* represented acting independently. *Competition* was added based on the tendency to innovate to beat out others. *Entrepreneurship* was identified as a third factor, which connected *Innovation* to a broader pair of constellations of traditional competencies (*Skill Set for Student*) and *Cooperation* (through *Identify Society Needs*). Some elements of these were noted to have some assessments (e.g., creativity, entrepreneurship, identifying others' needs), though a significant number of experts felt that the current state of the art was insufficient to model or predict meaningful innovation.

4) *Communication*: The center of Figure 2 is dominated by team and social competencies, which connect to *Communication* through *Trust*, *Cooperation*, and *Socio-Emotional Connectedness*. These were highlighted as key skills in society, which are assessed insufficiently, but which had many known assessment methodologies related to *Team Grades*, *Organization* (i.e., leadership/organizational theory), value-added to other members in a group (*Optimizing Others' Contributions*), and measures of *Empathy* and *Listening*. *Peer Assessment* and *Process/Patterns* (e.g., as analyzed using data mining) were considered central to measuring components underlying *Communication*.

5) *Subjects/What Society Learns:* The final major constellation of outcomes was related to assessing traditional academic and vocational competencies, and it could be said that the majority of this paper focuses on the bottom-left quadrant of Figure 2 (e.g., discussions on how to *Align to Performance*). While most of the items in this area are fairly typical (e.g., *Prerequisites*, *Learning Rates*), emphasis was also placed on *Metacognition* and related issues of both *Retaining Skills* and also *Forgetting/Pruning* outdated or irrelevant skills. *Literacy* and ongoing *Re-Defining of Literacy* were also explicitly highlighted, with the understanding that reading must expand to consider not just text comprehension, but also data graphs, digital literacy, and other as-yet-unknown critical understandings of symbols.

While this is clearly a very small slice of the space of outcomes that warrant assessment, there were a number of unanticipated implications. First, despite great advances, ITSs and educational data mining (EDM) can currently only meaningfully measure and interpret a small fraction of the outcomes of interest even among the Near-Term and Emergent/Intermediate categories. In some cases, this is because certain outcomes suggested are beyond the scope of an ITS (e.g., Gross Domestic Product growth). In other cases,

this is because the outcomes are inherently complex to assess, such as entrepreneurial skills. However, the majority of outcomes suggested by experts do have measures for assessing a learner's progress but are simply not measured by existing ITSs (e.g., identifying the needs of another person). For example, while certain ITSs provide learning experiences that may improve social interactions (e.g., team tutoring) or agency (e.g., raising self-efficacy), these are very seldom assessed meaningfully by the ITS and used for personalized adaptation for a student. As such, future ITSs might emphasize assessment of social skills (e.g., communication and group dynamics) and also on general adaptation/innovation factors (e.g., curiosity, agency, diversity of ideas).

Conclusions

Assessment has both benefits and pitfalls for the educational system. Achieving positive outcomes while avoiding unintended consequences of assessment requires careful planning not only at the level of individual assessments or assessment systems, but also at school, policy, state, or federal policy levels. A reasonable rule of thumb may be that no assessment should have stakes higher than we would trust its alignment to some set of real-world skills that learners actually need (e.g., for career or general life performance). This alignment does not (and often should not) imply a one-to-one relationship between assessments and real-life skills: education is intended to prepare learners for a broad range of experiences, many of which are unknown (e.g., who knows what careers programming might be useful for in 50 years?). With that said, there should be some chain of inferences that connect an assessment to expected real-life needs. For example, math problems to evaluate the costs of different credit card payment plans would be more ecologically valid than calculating how many washers and driers you could buy to spend exactly $1,245. This guideline inherently trends toward more anchored and complex problems, leading to assessments that capture the skills that society needs rather than skills that are highly reliable but have low ecological validity.

Similarly, it implies a sliding scale for aggregate performance by teachers and schools that is weighted based on such validity to avoid perverse incentives to overfit performance on trivial and low-fidelity tasks. This means evaluating both what we know and what we cannot yet measure to guide development of future assessments that can reduce our uncertainty about hard to measure but critically important skills. This has significant implications for ITSs as well, particularly in the context of machine learning. For example, reinforcement learning to optimize for posttest performance (a common metric) amplifies many of the risks stated here about teaching-to-the-test and, unlike a teacher, the machine will uncritically optimize for trivial tasks if told to do so. This implies the need for a more diverse set of rewards to optimize for, such as subjective assessments (e.g., peers, teachers).

Despite these pitfalls, we remain optimistic about the role of assessment in education and ITSs due to two trends. First of all, the set and types of skills assessed by digital systems is rapidly increasing. As more education moves into digital formats, we are starting to see traces of social skills, complex problem solving, and teamwork. Adapting educational technology systems to collect such information, and use it to assess such skills is the next great frontier in assessment. Second, we see increased focus on integrating evidence-based best practices with feedback for assessment. At this point, we have substantial research on how to effectively develop skills such as teamwork. While most measurements used in such research projects have yet to be applied in classrooms or commercial grade learning technology with solid statistical significance, the results of such research are increasingly informing practice.

Educational policy is slowly beginning to move from an unhealthy system of poorly proportional high-stakes outcomes for students, teachers, and schools, to one where assessment is ubiquitous, but high-stakes assessment is only one of many indicators of best practice. Qualitative feedback, through processes such as multiple instructors in classrooms providing peer feedback, is becoming increasingly common. A key pol-

icy issue will be integrating evidence-based best practices, qualitative assessment, and quantitative assessment into reasoned, thoughtful pedagogy decisions. These are socio-technical issues that have serious implications for large-scale use of learning technology in the future, where intelligent systems like the General Intelligent Framework for Tutoring (GIFT) will be responsible for both collecting and analyzing these data to help recommend learning resources and courses. However, such recommendations can only be successful if they align to broader educational and training goals. As such, educational assessment and artificial intelligence in education should share strong ties as both fields develop.

References

Anderson, J. R. & Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science, Vol. 5* (pp. 1–33). Mahwah, NJ: Erlbaum.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A. & Koedinger, K. (2008). Why students engage in" gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185.

Baker, R. S., de Carvalho, A. M. J. A., Raspat, J., Aleven, V., Corbett, A. T. & Koedinger, K. R. (2009). Educational software features that encourage and discourage "gaming the system". In *Proceedings of the 14th international conference on artificial intelligence in education* (pp. 475–482).

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2004). Working Inside the Black Box: Assessment for Learning in the Classroom. *Phi Delta Kappan*, *86*(1), 9–21.

Bloom, B. S. (1971). Mastery learning. *Mastery learning: Theory and practice*, 47–63.

Brusilovsky, P., Somyürek, S., Guerra, J., Hosseini, R. & Zadorozhny, V. (2015). The value of social: Comparing open student modeling and open social student modeling. In *User Modeling, Adaptation, and Personalization (UMAP) 2015.* (pp. 44–55). Springer.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A. & Koedinger, K. (2008). Why students engage in" gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185.

Chesler, N. C., Ruis, A. R., Collier, W., Swiecki, Z., Arastoopour, G. & Shaffer, D. W. (2015). A novel paradigm for engineering education: Virtual internships with individualized mentoring and assessment of engineering thinking. *Journal of biomechanical engineering*, *137*(2), 024701.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., ... & Ho, A. (2013). *Criteria for high-quality assessment*. Stanford, CA: Stanford Center for Opportunity Policy in Education

Embretson, S. E. & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Foltz, P. W. (2016). Advances in automated scoring of writing for performance assessments. In Y. Rosen, S. Ferrara & M. Mosharref (Eds.), *Handbook of Research on Tools for Real-World Skill Development*. Springer.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. Psychology Press.

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C. & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review*, *13*, 74–83.

Gulikers, J. T., Bastiaens, T. J. & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational technology research and development*, *52*(3), 67–86.

Hout, M. & Elliott, S. W. (Eds.). (2011). *Incentives and test-based accountability in education*. National Academies Press.

Kessler, A. M., Stein, M. K. & Schunn, C. D. (2015). Cognitive Demand of Model Tracing Tutor Tasks: Conceptualizing and Predicting How Deeply Students Engage. *Technology, Knowledge and Learning*, *20*(3), 317–337.

Kelvin, W. (1883). "Electrical Units of Measurement." Lecture. May 3, 1883.

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M. & Siemens, G. (2016). Towards Automated Content Analysis of Discussion Transcripts: A Cognitive Presence Case. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 15–24). New York, NY, USA: ACM. https://doi.org/10.1145/2883851.2883950

Kulik, C. L. C., Kulik, J. A. & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of educational research*, *60*(2), 265–299.

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity research journal*, *18*(1), 3–14.

Kessler, A. M., Stein, M. K. & Schunn, C. D. (2015). Cognitive Demand of Model Tracing Tutor Tasks: Conceptualizing and Predicting How Deeply Students Engage. *Technology, Knowledge and Learning*, *20*(3), 317–337.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. Educational Measurement: Issues and Practice, 23(4), 6–15.

Luo, H., Robinson, A. C. & Park, J. Y. (2014). Peer Grading in a MOOC: Reliability, Validity, and Perceived Effects. *Journal of Asynchronous Learning Networks*, *18*(2), 1–14.

Mills, C. N. & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9*(4), 287–304.

Mitros, P. (2015). Learnersourcing of complex assessments. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 317–320). ACM.

Mitros, P., Agarwal, A., Paruchuri, V. Assessment in Digital At-Scale Learning Environments by Piotr Mitros, Anant Agarwal, and Vik Paruchuri. In *ACM Ubiquity: MOOCs and Technology to Advance Learning and Learning Research*. April 2014.

Mehra, A. (2012). The JEE conundrum. *Current Science*, *103*(1), 29–36.

Nye, B. D., Goldberg, B. & Hu, X. (2015). Generalizing the Genres for ITS: Authoring Considerations for Representative Learning Tasks. *Design Recommendations for Intelligent Tutoring Systems: Volume 3: Authoring Tools and Expert Modeling Techniques* (pp. 47–64). US Army Research Laboratory

Nye, B. D. (2014). Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. In *International Journal of Artificial Intelligence in Education, 25* (2) 177–203.

Nye, B. D. (2016) Session on "Why do we assess?: What outcomes do we care about and how to measure them?" *Google EdFoo 2016*. Feb 19–21, 2016.

Schell, J., Lukoff, B. & Mazur, E. (2013). Catalyzing learner engagement using cutting-edge classroom response systems in higher education. *Cutting-edge Technologies in Higher Education*, *6*, 233–261.

Pavlik Jr, P., Bolster, T., Wu, S. M., Koedinger, K. & Macwhinney, B. (2008). Using optimally selected drill practice to train basic facts. In *Intelligent Tutoring Systems (ITS) 2008*, (pp. 593–602). Springer Berlin Heidelberg.

Rosen, Y., Ferrara, S. & Mosharref, M. (2016). *Handbook of research on technology tools for real-world skill development*. IGI Global. Hershey, PA.

Segedy, J. R., Kinnebrew, J. S., Goldberg, B. S., Sottilare, R. A. & Biswas, G. (2015). Using GIFT to Model and Support Students' Metacognition in the UrbanSim Open-Ended Learning Environment. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym3)* (p. 13–22).

Thissen-Roe, A., Hunt, E. & Minstrell, J. (2004). The DIAGNOSER project: Combining assessment and learning. Behavior research methods, instruments & computers, 36(2), 234–240.

Tang, S., Gogel, H., McBride, E. & Pardos, Z. A. (2015). Desirable Difficulty and Other Predictors of Effective Item Orderings. *International Educational Data Mining Society*.

Wiggins, G. (1990). The Case for Authentic Assessment. ERIC Digest.

CHAPTER 17 – Assessment of Forgetting

Philip I. Pavlik Jr.[1], Jaclyn K. Maass[1], and Jong W. Kim[2]
University of Memphis[1] , US Army Research Laboratory[2]

Introduction

Most efforts to educate individuals in a domain conclude with efforts to assess the effectiveness of the educational activities. During education research, the assessment activity tends to focus on the performance of the students and what that reflects on the quality of the educational intervention. After a quality educational intervention has been deployed, the focus of assessment turns to the individual students. In both cases, there are multiple considerations in collecting, analyzing, and interpreting the results of assessments.

Types of Knowledge to be Assessed

Starting with the assessment itself is the question of what to assess. This depends directly on the goals of the assessment. If the assessment is designed to measure what the student learned in the exact same form they learned it, then the test is probably best considered a test of recall of (factual) course content. If the assessment measures transfer to unpracticed problems that require some generalization of the material learned in the course, then the test is best considered a test of transfer (application). Of course, any assessment may measure both aspects of learning, which may be wise, since theories of learning have implied that simple factual knowledge underpins more complex or applied understandings (Bloom, Englehart, Furst, Hill & Krathwohl, 1956).

This distinction between factual and applied skills seems to be inherent in human brain processes, with evidence suggesting that learning may begin with episodic representations of events a student experiences (McRae & Jones, 2013). These episodic representations can be called upon when the student needs to retrieve knowledge of specific prior events. These simple prior events are typically not transferable, since they have no generality. When these prior events are repeated verbatim, the brain comes to represent this as a static commonality, AKA a fact. While this kind of memory is necessary for some performances, e.g., times tables, these contexts are limited, since factual representations need strong specific cues for retrieval. In other words, one only thinks “Paris” in reference to “the capital of France”, and this fact stands alone even if one knows nothing else about France.

Knowledge is likely generalized in two ways. First, it may become “semantic” in that it accumulates associations with multiple contexts or functions because it is repeated in a way that varies with each repetition. When exemplars represent an underlying type in this way, we can say they represent a category, and the learning of such categories is foundational to our current understanding of concepts. For example, our concept of the word “capital” is deeply connected to the many examples of capitals we have been familiar with, whether that is “D.C., Lansing, Paris, or London.” This conceptual knowledge supports the creation of on-the-fly “mental models” that allow us to represent and make decisions given more complex situations, for example, if we are in an unfamiliar capital, we might assume it has features of these other cities. Assessing such conceptual knowledge can be difficult, since it may be highly individualized. This makes it difficult to assess, since the materials used for assessment must be different from study materials to see whether the student has learned more than just the factual instances, but also the generalizable knowledge.

In addition to these types of semantic knowledge, there is production skill, which is perhaps most important in workplace environments. Production skill refers to some sort of specific function that needs to be assessed, such as the ability to hit a target or drive a complex piece of heavy equipment. These sorts of skills

range in complexity and in the amount of support they get from the episodic and semantic knowledge. Highly motor skills such as firing a weapon at a target may begin with semantic understanding of the proper breathing method and operation of the gun, but after long practice become generalized operational skills deployed with little if any thought about the tool being used. In psychological terms, they become automatized. Automatization requires large amounts of practice. In contrast, non-motor skills, such as tactical decision making also likely involve production knowledge, but in combination with a semantically supported episodically represented "situation model" that integrates their understanding of a complex situation with their production rules for action in such contexts.

Types of Assessment: Recognition, Recall, or Transfer

Each of these types of knowledge requires different considerations for assessment. This is because both the type of assessment tasks needed and the stability of the assessment vary as a function of type of knowledge being assessed. For example, often with text-based instruction, assessment types are defined based on the number of cues available to aid the learner in remembering the desired information. The formats of assessment for textual information range from recognition, e.g., true/false, multiple choice, to cued recall or fill-in-the-blank to free recall or essay prompts. The fewer cues available in the assessment question (i.e., free recall has the least amount of retrieval cues), the more rigorous it is as a test of memory.

There is also the consideration of how deeply to assess memory or learning. For example, when measuring reading comprehension, research by Kintsch and Van Dijk (1978) suggests two levels of text comprehension: a text-based model (comprehension specific to the information provided within the studied text) or situation-based model (incorporating knowledge of the information within the text to prior knowledge). If one is only interested in someone's memory for facts or information provided in a text, a relatively simple test of recognition or cued recall would likely be appropriate. If one intends to measure someone's situation-based model of text comprehension, questions that measure transfer would be ideal. Transfer refers to asking questions in a different context than they were originally provided (Barnett & Ceci, 2002). Transfer items are a much more difficult form of assessment, as many times people do not perform well in a situation or example that differs dramatically from the way in which they studied or practiced, which is sometimes referred to as transfer appropriate processing (Morris, Bransford & Franks, 1977).

However, not all learning is text-based. If one is learning a procedural or motor task (e.g., cooking, driving a car, etc.), it is also important that the assessment should be as close to end goal task as possible. This increases the accuracy of the assessment as a measure to see how the person would perform in the real world. Based on the notion of transfer appropriate processing (Morris, et al., 1977), being able to answer text-based multiple-choice questions about how to replace a bicycle chain is not often highly correlated with being able to perform that action. In other words, the choice of assessment should align both with the style or format of learning that has occurred, as well as with the ability, skill, or knowledge that one wants to assure the learner has gained.

Recency of Assessment Delivery and Interpretation

Just as the type of knowledge being assessed affects what type of assessment one uses, the recency of the assessment is also an important component to consider. We can talk of two types of recency: the recency of the assessment after the student learns the knowledge and the recency of the interpretation of the assessment results by stakeholders after it is completed. The former of these two types of recency is an especially important aspect to consider, as forgetting rates vary over time.

Recency of assessment considerations may be in inverse proportion to transferability of the knowledge. In other words, well-learned complex knowledge and production skills may be rather resistant to forgetting, while simpler memory skills are often forgotten quickly even after what appears to be mastery-level recall performance by the student. These differences in forgetting lead to the recency effect and come from many differences that are due to how the knowledge is learned. For example, from a cognitive neuroscience perspective, it may be that more stable production-type skills are encoded through many repetitions into the neocortex in a highly distributed manner that is resistant to forgetting, i.e., is more stable. In contrast, simple episodic memories, including recent verbal events, may be encoded in a relatively local compressed manner, likely involving the temporal lobe and hippocampus specifically, where they are quickly represented, but where the memories are very susceptible to interference and forgetting (McClelland, McNaughton & O'Reilly, 1995).

Recency effects may also come from the neural level, where researchers have described how more widely spaced stimulation results in stronger long-term potentiation of neural connectivity (Scharf et al., 2002; Wu, Deisseroth & Tsien, 2001). This may explain the well-known advantage of spaced practice, or correspondingly, the well-known disadvantage of massed practice or cramming. This effect may cause large difference in the stability of acquired knowledge, such that, particularly for verbal information, distribution of usage/practice over more time leads to longer-term learning. So for a course, particularly a course with a large amount of cumulative content where prior information must be maintained, distribution of the course over 16 one-hour sessions might be expected to result in much more long-term knowledge than two 8-hour sessions. However, assessors beware, if the two-session course has a test immediately afterward, they may show more learning than for the 16 one-hour sessions. This is because it is not the amount of learning that is improved by spaced practice; rather, it is the stability of what is learned. So, even though wide spacing results in more difficulty during practice, after a reasonable retention interval (perhaps 1 month) the distributed class might be expected to perform better when assessed. In neuroscience, this stabilization of memories is directly linked with long-term potentiation (LTP) of neural connections and formation of neural connectivity (Govindarajan, Israely, Huang & Tonegawa, 2011).

Recency effects make assessment even more difficult due to the changing rate of forgetting. This lack of constancy in the speed of forgetting may have been first described in terms of Jost's law, which says that given two memories of equal current strength (but of different ages), the older one will be forgotten more slowly (Simon, 1966). This implies that a power decay function (where a variable is raised to a constant negative power) may be an accurate way to represent forgetting. While this theory began as an observation about memory behavior, more recent work shows how averaging of multiple exponentials (such as might be expected if individual neurons had varying exponential decay rates) may produce a power law forgetting (Anderson & Tweney, 1997). Indeed, it has been suggested that in a healthy human, forgetting either becomes immeasurable or ceases after about 3 to 6 years (Bahrick, 1984).

Practical Assessment Guidelines

This discussion of forgetting leads to some practical guidelines for assessment. These guidelines assume that assessors desire a stable and accurate measure of the students' ability to perform in real job contexts. As the previous discussion suggests, interpreting these measures depends on the type of knowledge assessed and how students learned that knowledge. Therefore, the following guidelines are split into two categories.

Skills – Here a skill means a reflex-like complex behavior that students learn through extended practice and refinement. This includes skills such as driving, shooting, swimming, or archery, but also includes cognitive skills of the same nature, such as fluent language production or perceptual identification skills. These two categories of proficiencies are grouped together and probably share a basic perceptual/motor nature that is nonverbalizable and somewhat automatic as proficiency increases. One thing to note about

skills is that initial practice for a skill is often practice of knowledge, our second category. It is only after basic knowledge about a skill is learned that a student can "compile" that knowledge to form a skill and begin refining that skill (Taatgen & Lee, 2003). We can assess skills during learning, i.e., microgenetically, because skills are typically more stable and less likely to show decrement over time. For example, in a marksmanship class, measuring the shooting accuracy as student learns in the class, will be effective in measuring their long-term retention, since these production rule skills are only forgotten very slowly and are easily maintained. Despite this ability to do assessment while learning, skills of this sort may need varying practice if we care about the skill generalizing to similar situation, i.e., practice with shooting while standing up transferring to shooting while crouched (Judd, 1908). This well-established finding implies that for optimal transfer we should configure learning to vary each repetition during practice.

Knowledge – Here knowledge refers to facts and other concepts that can be verbalized. While such knowledge is often called declarative, this sort of knowledge may include images or pictures to the extent that they are not yet compiled for automatic access as skills. While skills are likely learned after many trials, knowledge may be learned up to 100% recall after only a single encounter. However, such knowledge is often very difficult to remember at a delay, particularly if the knowledge is disconnected from deeper meaning.

Because of this rapid forgetting, which has a power function form, the assessment of knowledge may require three measurements to accurately determine future proficiency. With only a single, one-time measurement, the stability of the knowledge cannot be determined. On the other hand, with two measurements, most usually the last measurement during practice and some posttest measurement after a delay, we can see how quickly information is being forgotten. But with just two points of measurement, the curvature of the forgetting function cannot be estimated, since two-point measurement only allows inference of forgetting if we expect forgetting to be linear with time. Since forgetting is not linear, but decreases with time, measurements at three time points are needed. A three-point measurement allows the assessor to estimate the rate at which forgetting is slowing. This can be done mathematically by fitting a power function to the data, or heuristically, by observing that the proportion forgotten in the first interval is much greater than the proportion forgotten in the second interval.

This difference in forgetting rates also reveals a distinct and serious problem with knowledge assessments relative to procedural skill assessments. Knowledge assessments can be gamed by the student who expects an upcoming exam and crams before the exam. Since declarative knowledge can be learned quickly, this strategy can be effective at maximizing short-term recall, but the exam score cannot be trusted to reflect long-term proficiency. While this crammed knowledge will be available in the short-term, as with any newly learned knowledge, it will be forgotten quickly. This problem is quite intractable, with the only clear solution being to give exams to students without warning. If a knowledge quiz is unexpected by the student, there is no opportunity to game the system, and assessment can be trusted to reflect the current state of long-term learning in the student.

Conclusion

In this chapter, we have discussed different types of knowledge and assessment. Particularly, our discussion asserts the necessity of an intelligent framework for assessment and its interpretation when it comes to forgetting and its assessment. Long-term focused proficiency may not be always necessary (e.g., when use of the knowledge and skill provides frequent practice), but it is needed for some types of knowledge and skills that are infrequently used but critical to respond to a certain situation (e.g., emergency response skills). A period of disuse of knowledge and skills can cause performance degradation. Assessment of forgetting should be employed to better maintain such types of knowledge and skills. It is, therefore, necessary to

provide an improved framework to assess stability and transferability of the acquired knowledge and skills in an unannounced and unobtrusive way.

This problem can be approached by using and extending an intelligent tutoring system (ITS). One of such systems is the Generalized Intelligent Framework for Tutoring (GIFT; Sottilare & Goldberg, 2012; Sottilare, Brawner, Sinatra & Johnston, 2017). GIFT can support learning and assessment of the knowledge types discussed earlier in this chapter (e.g., simple recognition, cued recall, transfer of knowledge). In the GIFT's modular construct, the pedagogical module manages courses that the student learns. The student is guided to go through four quadrants (rules, examples, recall, and practice) based on the component display theory (CDT; Merrill, 1983). In the recall quadrant, cued-recall and recognition can be assessed. Also, in the practice quadrant, the student performs the task in a simulated environment, in which the procedural skill can be assessed using a microgenetic approach.

For example, a marksmanship study in GIFT (Goldberg & Hoffman, 2015) affords a real-time assessment by gathering physiological states (i.e., breathing, heart rate variability), and performance (e.g., scores, movements, and accuracy), which are tested against the expert model in real time. In GIFT, a hierarchy of concepts being taught, which is implemented in domain module, can be also expanded to deal with the ontological representation of knowledge, and the microgenetic nature of procedural skills (e.g., tasks and subtasks, skills and subskills, or movements or submovements). A microgenetic approach to assess forgetting in an ITS will help us to better identify the learner state and support improved knowledge and skill proficiency.

A main challenge of this work is to support ITS-based optimal training schedules. The training proponent needs to decide retention needs (e.g., a decision on a training program that requires a long-term proficiency) by manipulating training regimens. That is, a massed or spaced training regimen can promote different rates of forgetting (or different rates of learning and retention). Currently, GIFT modules assess the learner state without considering the forgetting of knowledge. Thus, it is worth addressing the assessment of the different types of knowledge and skills in GIFT. By identifying the different forgetting rates for different types of knowledge in GIFT we may be able to determine the optimal training schedule in terms of retention needs, which will help to identify a way to achieve training proficiency.

References

Anderson, R. B. & Tweney, R. D. (1997). Artifactual power curves in forgetting. Memory & Cognition, 25(5), 724–730.

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. Journal of Experimental Psychology: General, 113(1), 1–29.

Barnett, S. M. & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. Psychological Bulletin, 128(4), 612–637.

Bloom, B., Englehart, M., Furst, E., Hill, W. & Krathwohl, D. (1956). Taxonomy of educational objectives: The classification of educational goals. London: Longmans, Green and Co. Ltd.

Goldberg, B. & Hoffman, M. (2015). Adaptive course flow and sequencing through the engine for management of adaptive pedagogy (EMAP). In B. Goldberg, R. Sottilare, A. Sinatra, K. Brawner & S. Ososky (Eds.), Proceedings of the AIED Workshop on Developing a Generalized Intelligent Framework for Tutoring (GIFT): Informing Design through a Community of Practice (Vol. 6, pp. 46–53). Madrid, Spain.

Govindarajan, A., Israely, I., Huang, S.-Y. & Tonegawa, S. (2011). The Dendritic Branch Is the Preferred Integrative Unit for Protein Synthesis-Dependent LTP. Neuron, 69(1), 132–146.

Judd, C. H. (1908). Special Training and General Intelligence. Education Review, 36, 28–42.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85(5), 363–394.

McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102(3), 419–437.
McRae, K. & Jones, M. N. (2013). Semantic Memory, The Oxford Handbook of Cognitive Psychology.
Merrill, M. D. (1983). Component display theory. In C. M. Reigeluth (Ed.), Instructional-design theories and models: An overview of their current status (pp. 282–333). Hillsdale, NJ: Lawrence Erlbaum Associates.
Morris, C. D., Bransford, J. D. & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behavior, 16(5), 519–533.
Scharf, M. T., Woo, N. H., Lattal, K. M., Young, J. Z., Nguyen, P. V. & Abel, T. (2002). Protein synthesis is required for the enhancement of long-term potentiation and long-term memory by spaced training. Journal of Neurophysiology, 87(6), 2770–2777.
Simon, H. A. (1966). A note on Jost's law and exponential forgetting. Psychometrika, 31(4), 505-506.
Sottilare, R. & Goldberg, B. (2012). Designing adaptive computer-based tutoring systems to accelerate learning and facilitate retention. Journal of Cognitive Technology, 17(1), 19–33.
Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017. DOI: 10.13140/RG.2.2.12941.54244.
Taatgen, N. A. & Lee, F. J. (2003). Production compilation: Simple mechanism to model complex skill acquisition. Human Factors, 45(1), 61–76.
Wu, G. Y., Deisseroth, K. & Tsien, R. W. (2001). Spaced stimuli stabilize MAPK pathway activation and its effects on dendritic morphology. Nature Neuroscience, 4(2), 151–158.

CHAPTER 18 – Validity Issues and Concerns for Technology-based Performance Assessments

Irvin R. Katz[1], Michelle M. LaMar[1], Randall Spain[2], Juan Diego Zapata-Rivera[1], Jo-Anne Baird[3], and Samuel Greiff[4]
Educational Testing Service[1], RTI International[2], Oxford University[3], University of Luxembourg[4]

Introduction

Advancements in technology have led to a revolution in assessments. No longer limited to the bubble-and-booklet approach of the 20th century, today's assessments may involve rich, interactive exercises, assess new and complex constructs, and automatically record and score evidence of skills. Still, the inferences drawn from technology-rich assessments must adhere to the same principles of good measurement as do traditional assessments. In other words, the inferences or claims made about test-takers (e.g., "this learner is ready to move on to the next unit") must be valid: having sufficient empirical, theoretical, and logical backing. The process of collecting evidence that supports a particular interpretation of assessment results, in the context of a particular use of those results, is called validation.

The purpose of this chapter is to introduce the concepts of assessment validity and validation evidence, with the hope that this material serves as a framework for understanding how validity theory and practice may help the research and development of intelligent tutoring systems (ITSs). Our focus is on technology-based assessment, but the reader should interpret "assessment" broadly to include any task – from multiple-choice questions to essays to complex, interactive performance tasks, such as solving algebraic equations, conducting scientific experiments using a simulation-based environment, discussing a story with a virtual peer, or playing an online game. As long as the tasks produce information that allows someone (e.g., parent, teacher, test-taker, administrator) or a system (e.g., an ITS) to make a decision about test-takers (students), we have a situation in which the concepts of assessment validity and validation can apply.

Assessments differ, of course. Assessments target different knowledge and skills, and may engage those knowledge and skills at different levels (e.g., a test of recalled facts, an essay to evaluate one's ability to convey a reasoned argument). Assessments also have different purposes or uses (e.g., to predict college success, to provide formative feedback to help guide instruction). While the specifics of validity and validation may change depending on these assessment factors (especially the intended use of the assessment; Kane, 2013), and this chapter only scratches the surface of a full accounting of validity issues, the general framework presented in this chapter should help researchers in making decisions about how to approach validity and validation in a variety of assessment circumstances.

While validity is a well-known concept as applied to assessment, validity is also vitally important to ITSs and other forms of technology-based learning environments that use student inputs to drive their learning experience. These technologies use scores from assessments, usually inherent in the ITS, to make inferences about student mastery and learning state and to guide instruction. Without an appropriate assessment as an integral part of the learning environment, ITSs cannot make accurate decisions about student competency or make optimal decisions about what to show a student next (i.e., instructional sequencing). An error in any of these inferences can significantly impact the accuracy and effectiveness of an ITS. For example, inaccurate interpretations of test scores or student performance during ITS activities can produce inaccurate estimates of student proficiency and mastery within the student model. As a result, the tutor may offer the wrong form of support or feedback or the wrong type of remedial content, or may place the students within the wrong instructional track. These actions not only reduce instructional efficacy, they may also cause the learner to become frustrated with the system, undermining the students' motivation. Just as a good human

tutor needs to know how to interpret the actions of students on learning activities or the scores from an assessment to provide optimal instruction, so too does a good tutor within an ITS.

In the next section, we discuss how advances in technology have led to new types of assessments that blur the distinction between learning and assessment tasks. We then describe a general framework for thinking about validity and validation with respect to technology-rich assessments and their ITS counterparts, including a discussion of the different types of evidence researchers might use in their validation efforts. Then we examine validity issues and concerns for three example technology-rich assessment techniques: assessments that provide practice and feedback on targeted topics (mastery-learning tutors), assessments involving simulated dialog with on-screen avatars (conversation-based assessment), and assessments in which performances are collected during students' ongoing interaction with technology environments (stealth assessment). We conclude with implications for the Generalized Intelligent Framework for Tutoring (GIFT).

Related Research

Technology-Rich Assessment, or When Did Assessments Get So Similar to ITSs?

Technology has a long history in the field of assessment. In the 1960s, technology first changed the nature of the scores that could be reported for traditional, dichotomously scored (e.g., multiple-choice and constructed response) paper-and-pencil tests because of advances in computation-intensive psychometric models. These models allowed for efficient estimation of abilities (and shorter test-taking time) through computer-adaptive testing (e.g., Lord, 1970). The movement toward performance assessment in the later decades of the 20th century brought automatic scoring of essays and, eventually, interactive performance assessments that allowed content and interactions not feasible without digital technology (e.g., viewing the inside of a volcano), and automated scoring that incorporated automatic evaluation of solution procedures. Current research expands the range of assessment contexts (games, conversations) and types of evidence collected (interactive action logs; see Katz & Gorin, 2016, for more details on this chronology of events).

Some assessments that incorporate new technology are large-scale assessments designed to provide a snapshot of, for example, the skills of students of a certain age in skills that are considered relevant for later success in life across a nation or even across several nations. A good example of such a large assessment is the Technology and Engineering Literacy test recently administered as part of the National Assessment of Educational Progress (National Center for Education Statistics, 2016). The target domain is new for a large-scale assessment, involving reasoning about and with technology, skills that pervade students' school, work, and lives in the 21st century. The assessment poses a series of extended tasks in which students might design a terrarium, troubleshoot a hand-pumped well, or investigate the usefulness for making a decision of Internet-presented and other information, all delivered through simulated environments tailored to the particular task. For instance, to troubleshoot the hand-pumped well, students "try out" the pump, observe components that appear to be working or not working, and decide on approaches to fixing it. Decisions indicated through mouse-clicks, movements of the pump's components, and more traditional selections (e.g., multiple-choice questions) combine to indicate better or worse trouble-shooting approaches, which ultimately produce a task score that is not only composed of the overall performance outcome but also incorporates behavioral patterns that students exhibited while working on the task. Such an interactive assessment task is conceptually not that different from the types of interactive tasks posed by many ITSs or other advanced learning environments.

Relevant to our focus in this chapter on validity, technology has improved the assessment process in ways that are central to the evidentiary argument of the assessment – that is, how an assessment provides evidence that allow test users (e.g., teachers, parents, test-takers) to draw conclusions about test-takers. Computers are now used to administer both traditional tasks, previously presented on paper, and new item formats and

content that are only feasible (e.g., safe and cost-effective) through computer administration. Via this computer administration, human behaviors and interactions with test content can be captured and stored, either for immediate or later processing. This increases the amount and type of evidence available to support drawing conclusions about test-takers based on their assessment performance. Sophisticated computational models and algorithms emerging from various fields, including mathematics, statistics, psychometrics, computational linguistics, natural language processing, and computer science offer innovative tools for both the modeling and scoring of the evidence that our computers can now capture. We can even use computers to administer more tailored, adaptive tests that are responsive to individuals' ability levels, interests, and engagement, all of which serve to improve the reliability and validity of our score interpretations. In fact, such tailored, adaptive assessments share many properties with tasks delivered through ITSs or other computer learning environments, blurring the distinction between learning and assessment (Bennett, 2015). It is this combination of learning and assessment that renders the question of validity fundamentally important for ITSs.

Validity and Validation

Assessment may be considered as a form of evidentiary argument, for which evidence is collected in support of a particular claim that is defined by the testing purpose. The quality of the assessment is defined by the degree to which the evidence that is collected can be interpreted as persuasive and compelling evidence of the intended claim (Gorin, 2012). Taking this viewpoint, assessments create observational opportunities that give rise to evidence of test-taker knowledge and skills, producing scores (or other descriptions) that allow the assessment stakeholders to apply appropriate tools to interpret that evidence.

The interpretation of assessment results is generally made in the context of a claim about the test-takers. A licensure exam makes the claim that test-takers who pass are qualified to practice the relevant profession. The Scholastic Aptitude Test (SAT) claims that students who score higher are more likely to be successful in their first year of college (higher first-year GPA). An end-of-chapter test claims that passing students have learned most of the material contained in the chapter. Michael Kane has argued that the unifying purpose of validity studies is to validate the claims that assessment developers make. To support those claims, a validity argument must be crafted (Kane, 2013).

The validity argument connects observed behaviors on the given tasks to the claims put forth by the assessment developers. Actual validity of assessment claims cannot be directly observed and can never be completely proven. However, the argument for that validity can be strengthened by both logical structure and collected empirical evidence that supports the validity of the interpretation. The shape of the validity argument and the amount and types of evidence needed to support it will depend strongly on the claims being made. Clearly, a licensure exam will require higher standards of validity than will the end-of-chapter test. For all assessments, however, some attention must be paid to the claims that are made and the evidence available to support those claims.

While many different types of evidence can be collected as part of a validity argument, the primary burden of validity evidence is to support the potentially weakest elements of the argument (Kane, 1992). Thus the evidence needed depends not only on the claims being made, but also on existing contrary evidence and plausible alternative explanations. Ordering becomes important here; the argument must be formulated and carefully examined first, and then the evidence to support (or refute) that argument must be sought.

Sources of Validity Evidence

Validation is the process of gathering and analyzing information to support our inferences. The results of the validation process present evidence that tell us what types of inferences can be made from scores obtained on an assessment. Five types of evidence are commonly examined to support the validity of an assessment. These include evidence of test content, response processes, internal structure, external structure (i.e., relation to other variables), and testing consequences (American Educational Research Association, American Psychological Association & National Council on Measurement in Education [AERA, APA & NCME], 2014).

Test content evidence is concerned with whether or not an assessment contains items and questions that adequately represent or cover the domain of interest. For example, a mathematics test that primarily contains addition problems would provide adequate evidence of a student's ability to add, but inadequate evidence of a student's ability to solve subtraction, multiplication, and division problems. As a result, the test may be less useful for making inferences about a student's overall mathematics ability. Evidence regarding test content is also concerned with item formatting, wording, scoring procedures, and the fairness of test items. Items and response formats that are inconsistent with the construct domain may be a source of irrelevant variance (i.e., affect scores in ways that are unrelated to the knowledge and skills being assessed), which can impact the validity of test score interpretations. Evidence about the appropriateness of content is most often determined on the basis of expert judgment (Webb, 2007). Usually subject-matter experts systematically judge the degree to which test items cover topics in the domain of interest and that scoring procedures are adequate. These ratings can be reviewed for consistency and reliability to determine how well a test aligns with the assessment plan. Unlike other sources of validity evidence, content evidence is primarily concerned with the design and construction of a test, rather than examining the empirical relation between test scores and other outcomes.

Evidence based on response processes is concerned with the degree of coverage between the constructs or knowledge components being assessed and how well a test's response format elicits and captures these processes. This form of validity evidence is similar to content validity in that it is concerned with the design of the test, and in particular whether the item response design facilitates or suppresses the types of cognitive processes expected for the item. For instance, a test that uses a series of multiple-choice question to assess high-order skills may be deficient in capturing the evidence needed to determine how well an individual can apply and integrate concepts and solve problems. A more appropriate response format may be a short or long essay in which students can construct an argument and provide concrete examples of how they would solve the problem, thus providing traces of their high-order thinking skills. Messick (1989) discusses several techniques that can be used to analyze the processes and strategies underlying task performance to guide the development of response formats or provide supporting validity evidence. One method is to conduct a protocol analysis in which students think aloud as they solve problems or describe retrospectively how they solved a problem (Ericsson & Simon, 1984). Another method is to ask students to provide a rationale for their answers or their way of responding to the item. The information gathered from these analyses can be used to ensure the test response format elicits the processes and reasoning they were intended to elicit, or help clarify the dimensions a test is measuring (Hamilton, Nassbaum & Snow, 1997). Ideally, these procedures would be used prior to developing a test to ensure the testing format captures the reasoning and knowledge components required to solve a problem or answer a question. The validity evidence drawn from the response process also extends to how raters view and interpret constructed-response items (Lane, 1999). In this case, evidence is needed to ensure that raters are interpreting and using scoring criteria adequately when grading constructed response items and performance based tests.

The third type of evidence that we discuss is internal structure evidence. This form of evidence is described as "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p. 16). Because constructs are

not directly observable, test items and the assessment as a whole must be designed to sufficiently operationalize the underlying characteristics of the trait or concept. Sources of internal structure evidence can come from reliability metrics, item fit statistics, analyses of item distractors (the incorrect multiple-choice options), or exploratory or confirmatory factor analyses. The latter allow researchers to examine item correlations and factor loading patterns. Using this information, researchers can determine if items are loading on the intended constructs or if more complex patterns are emerging from the assessment. If more complex patterns emerge, then this may suggest a test's content or response format are not properly aligned with the intended construct(s) (Lane, 1999).

Evidence based on relations to other criterion variables refers to traditional forms of validity evidence such as those collected through concurrent or predictive validity studies. It also extends to establishing evidence of convergent or discriminant validity (AERA et al., 2014, p. 16). Convergent evidence is provided by showing that scores from an assessment correlate with similar measures that assess the same construct. Discriminant evidence is provided by showing that scores are less related to assessments that measure different or unrelated constructs. Criterion-related evidence is demonstrated by showing relations between test scores and performance on a criterion measure or different student outcomes (such as training performance, job performance, grade point average, or another outcome of interest). For instance, in the context of employee selection an official may want to determine whether scores on a particular assessment, such as a work sample test, predict whether a candidate will perform well on-the-job. Ideally, high scores on the assessment activity should suggest high performance in the transfer domain. If test scores are correlated with relevant job criteria, then the official can begin to draw inferences about an individual's likely future job performance.

Gathering criterion-related evidence works particularly well in cases in which a good (e.g., appropriate, interpretable, useful) criterion score is readily available. If test scores are used to predict future performance on-the-job or in the classroom, and objective measures of performance are readily available, then correlation and regression concepts can be used to forecast how changes in test scores correspond to changes in the criterion (job performance ratings, course grades, etc.). The limitation of criterion-related evidence is that in some cases it can be difficult to develop or locate an acceptable criterion measure. In other cases, it can be difficult to implement a criterion that is better than the test itself or even to determine what an appropriate criterion is (Kane, 2013).

The fifth source of validity evidence is evidence based on consequences of testing. This source of evidence is concerned with intended and unintended consequences of testing and whether the benefits of using a test score have been realized. Evidence based on consequences of testing may include examining adverse impact (i.e., does a test unintentionally discriminate against a subgroup or population), evaluating the effects of testing on instruction (does testing improve instruction or does it introduce unintended changes or consequences), or examining how testing impacts other outcomes of interest (e.g., course pass rate; final grades, career opportunities). The AERA et al. (2014) note that gathering this evidence is particularly important when it can be used to determine weaknesses of a test or identify sources of construct contamination or deficiency.

What Validity is Not: Common Myths

One common myth about validity is that it is a property of the test. That is, a test is either valid or invalid, and this outcome is based on available evidence. However, validity is neither a single number nor a single argument, but an inference from all available sources (Guion & Gibson, 1988). Although it is true that assessments must demonstrate certain psychometric properties to maintain valid interpretations, tests themselves are not validated. Rather, it is the inferences from test scores that are validated (Kane, 2013). Accordingly, there can be as many validations as test usages; each usage or inferences requires sufficient

evidence to form an argument-based approach to validation. For instance, suppose we wanted to determine if performance on a verbal ability test predicted academic success. If we found test scores were correlated with relevant academic criteria, such as GPA, then we may draw certain inferences about the students who took the test. In this hypothetical example, we could conclude that test-takers might succeed in some future academic endeavor (or not, depending on their score). However, it would be inappropriate (i.e., not a valid inference) to use this same test to predict another outcome without evidence that this test also was useful for predicting this other outcome; perhaps one that was not at all related to the content within the test, such as mechanical aptitude. Thus, the test is not "invalid" generally, but use of the test and inferences drawn from it is supported by evidence for certain uses with certain test-taking populations and not others.

Another common myth is that there are different "kinds" of validity. Support for this widely held belief is rooted in many text books, and in historical discussions of validity, in which concepts such as content or criterion "validity" are described in succession without connecting them to an overarching model of evidence for the use of a test to make valid (supportable) inferences. This misconception might lead people to consider each "kind" of validity as operating separately in its own sandbox. However, consistent with the earlier discussion, many researchers consider validity to be a unified concept with these different "kinds" of validity serving as different sources of evidence (Messick, 1989). Each source provides an additional layer of evidence that can be used to strengthen the argument or case for validity.

Discussion

Validity within ITSs: An Example of Mastery Learning, Feedback, and Multiple Solution Attempts

ITSs often use a mastery-learning approach (Bloom, 1968) in which students practice a skill until they can show mastery of it (Corbett & Anderson, 1992; Heffernan & Heffernan, 2014). In traditional synchronized classroom instruction, a topic is studied for a fixed period of time after which students are evaluated on their understanding of the topic using an end-of-unit quiz. Independent of their quiz scores, all students are then moved on to the next topic. The idea behind mastery learning is to flip the equation so that rather than having fixed time and variable learning, students achieve fixed learning in variable time (Block & Burns, 1976). While this approach has been shown to achieve better outcomes than traditional instruction (Bloom, 1984), it is difficult to implement with a single teacher in a moderate to large classroom. Computerized instruction, on the other hand, more easily allows for individualized pacing and personalized instruction, making ITSs a natural medium for mastery-learning approaches.

Mastery learning, and ITS implementation of mastery learning in particular, presents interesting questions for assessment validity. Within the context of a mastery-learning tutor, a variety of claims can be entertained and the evaluation of assessment validity will depend upon which claim is being presented. The primary assessment claim centers on the mastery of the material being taught. The tutoring approach depends upon continuing instruction and practice until the topic has been mastered, so the assessment of mastery is key to proper implementation. As a categorical inference, determination of mastery is actually making two claims about students over time: 1) mastery-classified students have mastered the topic and are ready to move on to the next topic, and 2) non-mastery-classified students need more practice and/or instruction. In addition to the mastery classification required to make mastery-learning work, mastery learning can also be used as an assessment, making claims about differences in student ability. In the original formulation of mastery learning, it was proposed that differences in ability would manifest as differences in the time it took to achieve mastery (Block & Burns, 1976). Thus as an assessment, the methodology puts forth the claim that differences in ability within a particular domain can be distinguished by time-to-mastery

measures. A final claim that might be made, especially when evaluating ITSs compared to traditional instruction, is that students are learning when using the tutor. The claim of learning is different from the claim of knowing as it involves a change in state from not-knowing to knowing, and both states need to be correctly measured to achieve a valid inference that growth has occurred.

When viewed as an assessment, most computerized mastery-learning systems are similar in form to traditional assessments in that they present a series of problems to the student and collect and score student responses. A few critical differences, however, challenge the traditional assessment mechanics and therefore add complexity to issues of assessment validity. First, a mastery-learning system, such as an ITS, provides feedback or hints to the student when they are unsuccessful at a problem and the type and number of hints are not the same for all students. This type of scaffolding aids learning, but complicates assessment as it becomes unclear how much the student is relying upon the hints versus actually improving in performance. Similarly, mastery-learning systems often allow repeated attempts on the same problem, presenting questions about how to score a correct answer on a second or third try. Put differently, in this blending of learning and assessment experience, students at different levels are presented with different tests (e.g., some receive hints, others do not) making it difficult to derive scores that are comparable across individuals.

Thus for assessment validity, one of the primary challenges of scaffolding and repeated attempts lies in the question of how dependent the student is on the aid they are receiving from the system. If an item is presented as a fixed-choice, for example, the simple information that one choice is incorrect can greatly increase the probability of correctly answering the item on a second attempt, even without any learning taking place. In the case that there is more informative feedback, the student may be successful in later attempts due to context-specific learning, which may or may not generalize to other problems that require the same skill. It is worth noting that the features of scaffolding and repeat-attempts are also found in educational games (Shute, 2011) and dynamic assessments (Sternberg & Grigorenko, 2002). Thus the validity challenges produced by them are relevant to a wider array of educational instruments and are increasingly present in technology-based instruction and assessment.

To illustrate how validity arguments and validity evidence can be evaluated for these types of assessments, we examine the mastery-learning claim and outline the key elements of the validity argument, potential threats to validity, and what validity evidence might be gathered to support the argument. The validity argument for student mastery might look like this: 1) students are able to correctly respond to most of these items; 2) correctly responding to items indicates understanding of the specific topic represented in the item; 3) these items are representative of the topic being taught; 4) therefore, these students have mastered the topic being taught. Threats to validity can then be identified in the places where the argument might break down. How many items are required to make a mastery determination? How many must be correctly answered? Does a correct response imply understanding? Are the items representative and sufficiently covering the topic in question? Does a display of mastery on these items in this context imply mastery of the topic in other settings and/or times?

To address some of these questions, psychometric modeling can be useful. A common model used in mastery-learning situations is Bayesian knowledge tracing (BKT), in which the performance of the student over a series of presumed equivalent items is tracked as a time series to determine at what point in the series the student has achieved a high probability of having mastered the required skills (Corbett & Anderson, 1994). Validity evidence produced by BKT generally focuses on prediction: how well can the model predict the student's performance on the next item? If the model has correctly inferred the skill mastery profile of a given student, the prediction should fall within a reasonable confidence interval determined by the parameters of the model (such as the so-called slipping and guessing parameters). Here correct predication provides evidence of internal consistency of the measurement instrument as it shows that performance on past questions correlates highly with performance on future questions, suggesting both reliability and sufficient modeling of the content dimensionality. A validity threat remains, however, that the scaffolding provided

by the tutor is inflating the number of correct responses, making students appear to have mastered the topic before they actually have. An extension to BKT that incorporates scaffolding into the model has been developed and may alleviate the problem produced by different levels of feedback given to different students (Sao Pedro, Baker & Gobert, 2013). None of these models deal with the questions of content, however, and a careful review of the items presented and their mapping to the domain being taught will provide important content validity evidence toward the mastery argument. Further evidence to support a validity argument for skill mastery should include performance on appropriate transfer tasks and performance on similar tasks but after a delay period during which the immediacy of the provided scaffolding would be mitigated. Once all the evidence has been gathered, the validity of the inference, that the tutor is able to correctly diagnose mastery, can be either supported or refuted.

Validity Issues Surrounding Specific Technology-Based Assessment Techniques

Research on new types of performance assessments pose challenges for validation because research is done to explore general assessment capabilities rather than specific assessments. Validity is a characteristic of an inference that one makes based on assessment results (Kane, 2013), such that a test-taker has sufficient mastery of a topic (an accountability usage), needs a specific type of remediation (a type of formative assessment), or is ready to enter college (a prediction of future potential). Challenges for validation arise because without a specific context – a specific population, usage, or full targeted construct – a full validity argument would be premature. Nevertheless, validation efforts can begin just by collecting evidence regarding key validity claims of the assessment technology approach or threats to validity arising from the approach. That is, just as validation efforts for a mastery-learning assessment, as outlined earlier, might focus on the weakest portion of a validity argument, so too might validation efforts for an assessment technology focus on just those claims and threats to validity – the presumed weakest points of a validity argument for any particular assessment that would use the new technology.

In the next sections, we discuss two general assessment technologies. These technologies – conversation-based assessment and stealth assessment – have been used to create prototype assessments that might have specific uses. Nonetheless, these technologies raise validity concerns that likely span specific assessment uses. In each section, we first describe the technology and give examples of its use, then discuss the key claims or threats to validity posed by the technology. When possible, we provide examples of validation efforts, namely, empirical studies that examine the extent to which the threats to validity affect the interpretation of assessment results using some of the five types of validity evidence described earlier.

The different sources of validity evidence described earlier provide the backing for the claims within a validity argument. In a similar way, when we investigate claims of an assessment or possible threats to validity implied by new assessment technology outside of a particular assessment use, these sources of validity evidence might help to direct validation efforts. Knowing the different possible sources of validity evidence might help researchers to recognize possible validity threats (or assumed claims) that might not have recognized originally.

Conversation-Based Assessment (CBA)

Dialogue-based systems have been used in the area of ITSs to support student learning (Graesser, Person, Harter & Tutoring Research Group, 2001). These dialogue systems engage students in natural, written, or spoken conversations about different aspects of the domain; analyze and extract information from these conversations; and use it to react appropriately based on the goals of the system. Conversation-based assessments (CBAs) involve students interacting with one or more virtual characters using natural language (i.e., spoken or typed responses) or predefined responses (e.g., menu-driven conversations). By carefully designing the overall "space" (or script) of a conversation (Zapata-Rivera, Jackson & Katz, 2015), CBAs

seek to collect and evaluate rich evidence about students' knowledge, skills and other attributes (e.g., additional evidence that may be difficult to obtain using traditional assessment approaches), provide test-takers with multiple opportunities to elaborate or demonstrate their knowledge/skills, and elicit explanations about decisions that students make while interacting with a task (e.g., a simulation scenario; Jackson & Zapata-Rivera, 2015). The CBA approach builds on advances in natural language processing technology (e.g., Dialogue and Speech Systems; Adamson, Dyke, Jang &, Rosé, 2014; Graesser et al., 2001; Graesser et al., 2004; Millis et al., 2011) as well as advances in technology-enhanced assessments (Bennett, Persky, Weiss & Jenkins, 2007; Clarke-Midura, Code, Dede, Mayrath & Zap, 2011; Quellmalz et al., 2011).

Figure 1 shows a screenshot of a CBA prototype used to measure both English language and mathematics skills. In the scene depicted in the figure, the student is interacting with two virtual classmates by taking turns at answering mathematics word problem questions. These questions involve understanding the problem, finding relevant information in tables, and applying the concept of ratio to find the answer. Virtual characters provide feedback and ask additional questions based on the student's responses. Other CBA prototypes have been developed and used to measure skills such as science inquiry (Zapata-Rivera et al., 2014), collaborative problem solving (L. Liu, Hao, von Davier, Kyllonen & Zapata-Rivera, 2016), language argumentation (Song, Sparks, Brantley, Oliveri & Zapata-Rivera, 2014), mathematical argumentation (Cayton-Hodges, Bauer, Bertling, Katz & Wylie, 2015), and English language skills (So, Zapata-Rivera, Cho, Luce & Battistini, 2015).

Figure 1. A screenshot of a CBA prototype designed to assess English language and mathematics skill.

Validity Issues and Concerns

In considering possible approaches to validation for CBAs, we may start by thinking about the possible threats to the validity of inferences drawn about test-takers based on their performance in a CBA. Of course, valid inferences cannot be separated from the assessment context (e.g., are we making a prediction of future performance? Are we recommending areas for remedial support?), but the general CBA approach of interactive dialogs with virtual characters suggest possible areas in need of validation efforts.

To what extent does the virtual interaction of a CBA distract from the construct of interest? The CBA environment might introduce construct-irrelevant variance, whereby knowledge and skills unrelated to those intended to be assessed affect test-takers' performance. For example, because of the dialogic nature of tasks, CBAs might involve more linguistic skills than a multiple-choice assessment. If an English language learner performs poorly on a CBA of scientific-inquiry skills it might be because that learner isn't able to express understanding via a dialog rather than due to a lack of science-inquiry skills.

To what extent does the CBA environment help us to assess unique knowledge and skills compared with more traditional forms of assessment? In other words, should we be using a CBA if a more traditional (and possibly more familiar and less time-consuming) assessment would suffice? The intended realistic interaction might encourage more meaningful preparatory instruction than is typically associated with preparation for multiple-choice tests (e.g., a focus on lower-level skills; Madaus, Russell & Higgins, 2009). For such an outcome to be meaningful, however, the CBA approach should provide at least some type of measurement benefit as well.

In the next sections, we discuss some of the validation research that has been conducted using particular CBA prototypes. To better connect the validation approaches to the earlier discussion of validity, we use **bold type** when we mention particular sources of validity evidence.

Scientific-inquiry Assessment

The Volcano Scenario is a CBA prototype that was designed to measure science-inquiry skills. In this scenario, students play the role of an apprentice to Dr. Garcia, a professional volcanologist. After learning about volcanoes (parts of a volcano, information about volcanic events, seismometers, seismic events, and alert levels), the student is asked to place seismometers to collect data on a volcano, annotate the data collected, and make a prediction about the volcano's alert level. The scenario includes several conversations with both Dr. Garcia and another apprentice (Art) about topics such as the quality of the annotated data, common misconceptions, and evidence supporting a volcano eruption prediction. A form of validity evidence related to the **response processes** elicited by the scenario was investigated through a small-scale study (N = 10 middle school students; Zapata-Rivera et al., 2014). The students indicated that they were able to complete the activity with minimal instruction, enjoyed the activity, felt that the virtual characters (Dr. Garcia and Art) understood their responses, and expressed that they would like to engage into similar conversation in the science classroom

Zapata-Rivera, Liu, Chen, Hao & von Davier (2016) investigated the internal structure and external relations (two types of validity evidence) of the Volcano CBA. Five hundred adults (via Amazon Mechanical Turk) completed the scenario, which resulted in four scores based on performance during the conversation and two scores based on multiple-choice questions embedded in the conversational scenario that were designed to assess similar scientific-inquiry skills. The participants also completed a separate multiple-choice assessment of general science knowledge. All of the scenario-generated measures correlated moderately with the general science assessment (correlations approximately 0.3), providing (albeit weak) evidence that the conversation-based assessment taps the construct of interest. This result provides a type of **criterion-related** evidence, namely, convergent evidence of relation with a predicted criterion measure, although stronger evidence might have been obtained by using discriminant measures as well. However, while the conversation-generated scores correlated with each other (rs about 0.7), they did not correlate with the scores from the embedded multiple-choice questions. These results suggest that the conversational dialogs provide useful assessment beyond what might be possible with multiple choice, a type of evidence based on the **internal structure** of the assessment.

Validation efforts continue both for this CBA and another, parallel, CBA that seeks to assess the same knowledge and skills using a scenario related to weather. Sparks, Andrews, Zapata-Rivera, Lehman &

James (2016) demonstrated that the virtual characters might inadvertently introduce irrelevant variance. For example, when asked whether they agreed or disagreed with a virtual peer's answer, middle school participants (N = 145) appeared unwilling to explicitly disagree with the peer in front of Dr. Garcia despite the participants demonstrating elsewhere that they understood that the peer's response was incorrect. This type of evidence related to students' **response processes** suggests that such situations should be avoided when designing future CBA situations.

English Language Assessment

Scenario-based tasks such as those in a CBA might be particularly appropriate for assessing multiple language skills (listening, reading, writing, and speaking). In a CBA prototype designed around these ideas, students interact and complete tasks with a virtual teacher and two virtual classmates in several familiar situations, such as listening to directions from a teacher and reporting them to a classmate, engaging in classroom conversations with classmates, and telling a classmate about the rules in the library based on a posted sign (So et al., 2015). To mitigate the potential validity threat of having non-native English speakers type all their responses, the prototype includes opportunities for students to respond verbally, using automated speech recognition (Evanini et al., 2014).

Validation efforts are ongoing, but small-scale studies ($N \sim 20$) provide some preliminary evidence for the use of this CBA to assess English language skills. For example, So et al. (2015) investigated the prototype's use by English language learners in grades 3–5, who represented a range of native languages and different levels of English proficiency. Students' performance on the CBA matched their ability as judged by their teachers, with the top three students outperforming the bottom six students, a type of **criterion-related** evidence (convergent; the teachers' judgment serving as the criterion). The students reported enjoying talking with the virtual characters and felt that the characters understood them. Similar results were obtained by Evanini et al. (2014) in a study that focused on the performance of the automated speech recognition system of this prototype. The researchers demonstrated that the prototype's categorization of student responses was robust relative to errors made by the speech engine. Together, these results suggest that students were both able (by virtue of the system working) and willing (via their engagement) to interact with the CBA in an intended way to allow for assessment of their language skills, a type of **response process** evidence.

Stealth Assessment

The original idea of stealth assessment refers to embedding a performance assessment within a video game (Ventura & Shute, 2013), sometimes described as "evidence-based assessments that are woven directly and invisibly into the fabric of the gaming environment" (Shute, Leighton, Jang & Chu, 2016, pg. 52). However, the notion of assessing people as they are engaged in an activity is not new. In one sense, we are all doing this on an everyday level; making assessments about one another without applying explicit assessment tasks. Indeed some of the assessments we make have significant consequences. As Foucault (1975) wrote,

> The judges of normality are everywhere. We are in the society of the teacher-judge, the doctor-judge, the educator judge, the "social worker"-judge; it is on them that the universal reign of the normative is based; and each individual, wherever he may find himself, subjects to it his body, his gestures, his behavior, his aptitudes, his achievement. (pg. 304)

Yet, many of these judgments are of questionable validity – borne of bias and reaffirming current societal structures and practices. Notwithstanding, the electronic era offers greater opportunities for us to collect data and make inferences about people from such stealth assessment data.

The promise of stealth assessment is that during an activity, such as a computer game, "students naturally produce rich sequences of actions while performing complex tasks, drawing on the very skills or competencies that we want to assess" (Shute, Leighton et al., 2016, pg. 52). Without disrupting flow, data could be collected on multifaceted, dynamic performances that are as close to the behaviors, knowledge, and skills that are needed for the real-life performances that we want the students to demonstrate (Ke & Shute, 2015). On a more general level, stealth assessments could also be used to collect data beyond traditional measures of learning to include emotional responses to learning interactions and social skills (DeRosier, Craig & Sanchez, 2012). Thus, stealth assessments, with the affordances of computer-based systems, could provide a better platform than paper-based tests to allow students to demonstrate what they know and can do.

To achieve these promises requires careful validation efforts surrounding any application of stealth assessment.

Validity Issues and Concerns

Almost by definition, during a stealth assessment, learners (or test-takers) might not know that their performances are being monitored or that assessment data derived from their performances (such as scores delivered to a teacher or other assessment user) will be used to draw inferences about their knowledge and skills. Because people may perform differently when they know they are being assessed, designers of a stealth assessment must consider the ways that their implementation of this assessment technology (i.e., the particular assessment usage) allows for valid inferences to be drawn about test-takers. Additionally, threats to validity must be considered, such as the possible ways that the stealth assessment may inhibit valid inferences or raise concerns about fairness (e.g., making consequential decisions about test-takers based on data they did not know was being collected). Validation plans – empirical data collected – should be designed both to test hypothesized claims and gather evidence that disconfirms potential threats.

For example, high-stakes assessments have been heavily criticized for the effects they have upon teaching and learning, especially for inducing a narrow curriculum, enabling shallow learning, and drilling instructional practices (Madaus et al., 2009). Additionally, students might be too anxious under high-stakes testing conditions to show what they know and can do. A stealth assessment might provide a more natural, less anxiety-provoking method of assessment. Thus, validation efforts might focus on evidence related to the **consequences** of stealth assessment. For example, to what extent do stealth assessments have the purported benefits on a curriculum or on students' levels of test anxiety relative to high-stakes testing, and are there any unintended, less beneficial consequences?

At the same time, not knowing that one is being assessed might introduce *construct-irrelevant variance* (Haladyna & Downing, 2004), namely, performances that obscure inferences about what a student knows or can do relative to the target domain (construct) of an assessment. Learners' schemas for assessment include individual assessments taken in silence in an examination hall, with a proctor present (Richardson et al., 2002). When these schemas are not cued, we might not get the best performances from people and therefore draw the wrong conclusions about their capabilities and/or the rank order of the students. Students' motivation to perform well is directly related to how well their test scores reflect their knowledge and skills (O. L. Liu, Bridgeman & Adler, 2012). Additionally, stealth assessments might be embedded with learning activities. However, during learning, students might choose to try different techniques, just to experience and explore them. They might even try things in the knowledge that they will fail, or simply find out that they failed but learn from these mistakes. Making mistakes is important for learning, but complicates drawing valid inferences based on assessment results.

Another potential validity concern relates to the generalizability of stealth assessment performance. Games and simulations are invariably "small-world" variants on the real-world behaviors that we wish to encourage in learners. Just as with traditional testing, we are usually assessing only a small sample of the performance of interest. Like last century's complaint that some people were only book learners but could not apply their learning to anything practical, we must be careful that this century's learners can apply their learning beyond the confines of the game. We have to help learners to recognize how to transfer their learning, recognize situations in which it will be relevant, and map the structural features of the games' problems onto those that they will encounter beyond the game. Validation efforts should include investigating that the scores generated by the stealth assessments are generalizable to real-world performance, which might be investigated using measures hypothesized to reflect real-world performance (i.e., **criterion-related** evidence) or evidence of **response processes**. For the latter, we might investigate the extent to which test-takers' response processes on the assessment correspond with processes elicited by real-world (or real-world-like) tasks (see Snow & Katz, 2010, for an example of this latter approach for a [non-stealth] performance assessment).

Along these lines of thinking, Shute, Wang, Greiff, Zhao, and Moore (2016) report an example of how problem solving as an imminent 21st century skill can be captured through stealth assessment. In their study, 55 seventh grade students worked on a game that was a slightly modified version of the well-known *Plants vs. Zombies*™ *2* game. On the basis of a theoretically derived problem-solving model as an indicator of construct validity, several in-game measures were derived with the aim of extracting students' problem-solving skills. Importantly, the scores derived in the stealth assessment were related to external measures of problem solving indicating, to the extent possible, **criterion-related** evidence of the underlying stealth assessment in a game-like environment. For example, using tools efficiently and effectively correlated with a test of IQ as measure of simple problem solving ($r = 0.40$, $p < 0.01$ with Raven's Matrices) and a test of complex problem solving skill ($r = 0.41$, $p < 0.01$ with MicroDYN). Although these were modest correlations and questions remain about the extent of coverage of complex problem solving in *Plants vs. Zombies 2*, this research is a useful step forward (Shute et al., 2016). In another study, correlations were found between students' physics knowledge from a traditional test and their "gold trophies" in the game *Newton's Playground*™ (correlations between 0.22 and 0.40). Some very practical recommendations for stealth assessment designers were given by Wang, Shute, and Moore (2015), including that an external criterion measure should be identified in advance, so that designers can check whether their scores do in fact correlate meaningfully with the constructs of interest.

Stealth assessment has been discussed in relation to gaming in the literature to date, though its conceptual connection with big data and data mining or learning analytics is also apparent. The term surely applies when smart watches and bathroom scales send data to smart phones and the data are collated with health club attendance. These combined data might then be sent to the health insurance provider, affecting premiums. Normal ethical rules of transparency of the process, data privacy, and consent have not always been applied under these new modes of assessment (Prinsloo & Slade, 2013). Given that the outcomes of stealth assessments might have serious **consequences** for individuals, these ethical issues need to be considered, and validation efforts should evaluate any particular application of stealth assessment. As applications continue to be developed, there is no doubt that this area will attract attention with regard to protocols and legal requirements for ethical behavior on the part of the assessors.

Recommendations and Future Research

The application of technology to education in the form of ITSs, adaptive assessments, and game-based assessments has produced considerable excitement, extending the hope for improving both learning and the student experience of education. As educators, however, we must take care that what we claim as effective instruction really is effective and when we claim to be able to measure student understanding, those

measures allow for valid inferences to be made about test-takers. The study of assessment validity has a long history in education and its principles can be applied to these new forms of assessment. This chapter has introduced several key validity concepts and illustrated their use in several technology-rich contexts.

As noted previously, ITSs rely on assessments to make appropriate decisions about student knowledge states, instructional support and feedback (micro-adaptive strategies), and instructional sequencing (macro-adaptive strategies). Providing sources of validity evidence is paramount in establishing the benefits of ITSs. In this chapter, we have discussed different sources of validity evidence and provided examples of how researchers can gather evidence to support the validity argument for ITS assessments. As a research platform, GIFT can support the design and implementation of ITSs that collect different types of evidence to validate their use by providing tools and guidance on the types of studies that can be carried out. The GIFT community can be a good mechanism for disseminating these ideas. In particular, formalizing the notion of evidence (Zapata-Rivera, Brawner, Jackson & Katz, this volume) can facilitate the implementation of some of the ideas presented in this chapter.

References

Adamson, D., Dyke, G., Jang, H. J. & Rosé, C. P. (2014). Toward an agile approach to adapting dynamic collaboration support to student needs. *International Journal of Artificial Intelligence in Education, 24*, 91–121.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1), 370–407.

Bennett, R. E., Persky, H., Weiss, A. R. & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP technology based assessment project* (NCES Report No. 2007–466). Washington, DC: National Center for Education Statistics.

Block, J. H. & Burns, R. B. (1976). Mastery learning. *Review of Research in Education*, *4*, 3–49. https://doi.org/10.2307/1167112.

Bloom, B. S. (1968). *Learning for mastery. Instruction and curriculum*. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. (Reprinted from *Evaluation Comment, 1*(2)).

Bloom, B. S. (1984). The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.2307/1175554.

Cayton-Hodges, G., Bauer, M. I., Bertling, M., Katz, I. R. & Wylie, E. C. (2015, April). *Assessing mathematical argumentation and algebraic reasoning through automated conversations.* Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Clarke-Midura, J., Code, J., Dede, C., Mayrath, M. & Zap, N. (2011). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In M. C. Mayrath, J. Clarke-Midura & D. Robinson (Eds.). *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 125–147). Charlotte, NC: Information Age.

Corbett, A. T. & Anderson, J. R. (1992). Student modeling and mastery learning in a computer-based programming tutor. In C. Frasson, G. Gauthier & G. I. McCalla (Eds.), *International conference on intelligent tutoring systems* (pp. 413–420). Berlin, Germany: Springer. Retrieved from http://link.springer.com/chapter/10.1007/3-540-55606-0_49.

Corbett, A. T. & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

DeRosier, M. E., Craig, A. B. & Sanchez, R. P. (2012). ZooU: A stealth approach to social skills assessment in schools. *Advances in Human-Computer Interaction*, 1–7. doi:10.1155/2012/654791.

Ericsson, K. A. & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Evanini, K., So, Y., Tao, J., Zapata, D., Luce, C., Battistini, L. & Wang, X. (2014). Performance of a trialogue-based prototype system for English language assessment for young learners. *Proceedings of the Interspeech Workshop on Child Computer Interaction* (WOCCI 2014), Singapore, September 19, 2014.

Foucault, M. (1975). *Discipline and punish: The birth of the prison*. New York, NY: Random House. Retrieved from: https://zulfahmed.files.wordpress.com/2013/12/disciplineandpunish.pdf.
Gorin, J. S. (2012). *Assessment as evidential reasoning.* White paper commissioned by The Gordon Commission on the Future of Educational Assessment. Retrieved from http://gordoncommission.org/rsc/pdfs/gorin_assessment_evidential_reasoning.pdf.
Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A. M. & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments & Computers, 36,* 180–193.
Graesser, A. C., Person, N., Harter, D. & Tutoring Research Group. (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education, 12*, 257–279.
Guion, R. M. & Gibson, W. M. (1988). Personnel selection and placement. *Annual Review of Psychology, 39*(1), 349–374.
Haladyna, T. M. & Downing. S. M. (2004). Construct-irrelevant variance in high stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.
Hamilton, L. S., Nussbaum, E. M. & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181–200.
Heffernan, N. T. & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education, 24*, 470–497.
Jackson, T. & Zapata-Rivera, D. (2015). Conversation-based assessment. *R&D Connections. No 25*. Princeton, NJ: Educational Testing Service.
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.
Katz, I. R. & Gorin, J. S. (2016). Computerising assessment: Impacts on education stakeholders. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 472–489). New York, NY: Routledge.
Ke, F. & Shute, V. (2015). Design of game-based stealth assessment and learning support. In C. Sebastian Loh, Y. Sheng & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment and improvement* (pp. 301–318). New York, NY: Springer International.
Lane, S. (1999). *Validity evidence for assessments*. Edward F. Reidy Interactive Lecture Series, The National Center for the Improvement of Educational Assessment, Providence, RI. Retrieved from: http://nciea.org/publications/ValidityEvidence_Lane99.pdf.
Liu, O. L., Bridgeman, B. & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*, 352–362.
Liu, L., Hao, J., von Davier, A. A., Kyllonen, P. & Zapata-Rivera, D. (2016). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development* (pp. 344–359). Hershey, PA: IGI-Global.
Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing and guidance* (pp. 139–183). New York, NY: Harper & Row.
Madaus, G., Russell, M. & Higgins, J. (2009). *The paradoxes of high stakes testing. How they affect students, their parents, teachers, principals, schools and society*. Charlotte, NC: Information Age Publishing.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. C. & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou & J. Lakhmi (Eds.), *Serious games and edutainment applications* (pp. 169–196). London, England: Springer-Verlag.
National Center for Education Statistics (2016). *The nation's report card: Technology and engineering literacy.* Washington, DC: Author. Retrieved from: https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2016119.
Prinsloo, P. & Slade, S. (2013). An evaluation of policy frameworks for addressing ethical considerations in learning analytics. In *LAK '13. Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 240–244). New York, NY: Association for Computing Machinery.
Quellmalz, E. S., Timms, M. J., Buckley, B. C., Davenport, J., Loveland, M. & Silberglitt, M. D. (2011). 21st century dynamic assessment. In M. C. Mayrath, J. Clarke-Midura & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 55–90). Charlotte, NC: Information Age.

Richardson, M., Baird, J., Ridgway, J., Ripley, M., Shorrocks-Taylor, D. & Swan, M. (2002). Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving. *Computers in Human Behavior, 18*, 633–649.

Sao Pedro, M., Baker, R. & Gobert, J. (2013). Incorporating scaffolding and tutor context into Bayesian knowledge tracing to predict inquiry skill acquisition. In S.K. D'Mello, R.A. Calvo & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 185–192). Memphis, TN: International Data Mining Society.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–523). Charlotte, NC: Information Age Publishers.

Shute, V. J., Leighton, J. P., Jang, E. E. & Chu, M.-W. (2016) Advances in the science of assessment, *Educational Assessment, 21*, 34–59, doi: 10.1080/10627197.2015.1127752.

Shute, V. J., Wang, L., Greiff, S., Zhao, W. & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117.

Snow, E. & Katz, I. R. (2010). Using cognitive interviews and student response processes to validate an interpretive argument for the ETS iSkills™ assessment. *Communications in Information Literacy, 39*, 99–127.

So, Y., Zapata-Rivera, D., Cho, Y., Luce, C. & Battistini, L. (2015). Using trialogues to measure English language skills. *Educational Technology and Society, 18*(2), 21–32.

Song, Y., Sparks, J., Brantley, W., Oliveri, M. & Zapata-Rivera, D. (2014, April). *Designing game activities to assess students' argumentation skills*. Paper presented at the meeting of the American Educational Research Association , Philadelphia, PA.

Sparks, J. R., Andrews, J. J., Zapata-Rivera, D., Lehman, B. & James, K. (2016*). Students' perceptions of collaborator expertise in science inquiry tasks using simulated conversations*. Paper submitted for presentation.

Sternberg, R. J. & Grigorenko, E. (2002). *Dynamic testing: The nature and measurement of learning potential*. New York, NY: Cambridge University Press.

Ventura, M. & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior, 29*, 2568–2572.

Wang, L., Shute, V. & Moore, G. R. (2015). Lessons learned and best practices of stealth assessment. *International Journal of Gaming and Computer-Mediated Simulations, 7*(4), 66–87.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards. *Applied Measurement in Education, 20ˆ, 7–25.*

Zapata-Rivera, D., Jackson, T. & Katz, I. R. (2015). Authoring conversation-based assessment scenarios. In R. A. Sottilare, A. C. Graesser, X. Hu & K. Brawner (Eds.), *Design recommendations for intelligent tutoring systems, Volume 3: Authoring tools and expert modeling techniques* (pp. 169–178). Orlando, FL: US Army Research Laboratory.

Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M. & Katz, I. R. (2014). Science inquiry skills using trialogues. In S. Trausan-Matu, K. Boyer, M. Crosby & K. Panourgia (Eds.), *Proceedings of the 12th International conference on Intelligence Tutoring Systems. Honolulu, HI, June 2014. Lecture notes in computer science* (Vol 8474, pp. 625–626). Cham, Switzerland: Springer International.

Zapata-Rivera, D., Liu, L., Chen, L., Hao, J. & von Davier, A. (2016). Assessing science inquiry skills in immersive, conversation-based systems. In B. K. Daniel (Ed.), *Big data and learning analytics in higher education* (pp. 237–252). Cham, Switzerland: Springer International. doi:10.1007/978-3-319-06520-5_14..

CHAPTER 19 – Toward Systematic Assessment of Human Performance Interventions in the US Army: An Assessment Process Framework

Kara L. Orvis[1], Jared T. Freeman[1], Jeffrey M. Beaubien[1], Clayton W. Burford[2], Joan H. Johnston[2], Lauren Reinerman-Jones[3], and Grace Teo[3]
Aptima, Inc. [1], US Army Research Laboratory[2], University of Florida, Institute for Simulation and Training[3]

Background

For over 100 years, the Army has relied on systematic assessments of human performance to help ensure force readiness. For example, upon America's entrance into World War I, the Army developed the Alpha and Beta tests to quickly and efficiently select potential recruits. By the mid-1960s, the Army adopted the Armed Forces Qualification Test (AFQT), which remains in use today, to place recruits into different Military Occupational Specialties (MOS) based on their unique profile of aptitudes. More recently, the Army has developed the Officer Evaluation Report (OER) system to manage its leadership pipeline; Mission Essential Task Lists (METLs) to assess individual and unit readiness at the conclusion of training; and the Functional Solutions Analysis (FSA) process to assess the effectiveness of prototype tools and technologies prior to deployment in the field.

While significant progress has been made in assessing various aspects of individual and unit performance, there is no unified, standardized measurement framework that can sustain progress over time. As a general rule, human performance-related data collection activities tend to be disparate and unorganized, thereby minimizing the possibility of reusing critical tools, techniques, and lessons learned. Additionally, there is no centralized mechanism for the aggregation of results across studies for use in meta-analysis. Assessments also rarely address longitudinal effects. Finally, a majority of current Army systems assume stereotyped human input that is geared toward an "average" soldier, rather than capitalizing on known variability in intelligence, personality, physical endurance, and related attributes. Taken together, these factors have decreased the Army's return on its investment in the science of human performance assessment.

In 2016, the US Army Research Laboratory (ARL) launched the Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness (UMMPIREE) project to design and build a framework that will "standardize the vocabulary on human performance assessment; promote the establishment of 'best fit' measurement in project- or context-specific assessments; and ultimately, increase confidence in the results of human performance experiments and interventions". UMMPIREE is a means to improve the assessment of individual and/or unit performance so that Army leaders can better use them in making decisions on topics such as personnel selection, training, promotion, performance augmentation, and related issues. Ultimately, the capabilities developed during the UMMPIREE project will be a point of departure for other applied research and development (R&D) efforts, thereby saving time and money, while at the same time facilitating rapid and reliable entry of assessment-related results into operational Army systems.

The purpose of this chapter is to present some of the ideas emerging from the UMMPIREE effort. Specifically to present a definition of assessment and a framework that depicts steps regarding how assessment should be accomplished. The framework is applied to assessing human performance in technologically-augmented learning environments such as intelligent tutoring systems (ITSs), specifically the Generalized Intelligent Framework for Tutoring (GIFT).

Defining Assessment

For the purpose of this chapter, we define assessment as an evaluation on some metric of the state, behaviors, or affect of an agent in a given context at a given point in time. The evaluation is typically a qualitative statement about the capability or readiness of the agent (e.g., *passing* a test, *mastering* a skill, or being *ready* for deployment). The agent can be a person, a team, or a team of teams, while the performance context is the setting in which the agent takes, or fails to take, various actions. Generally speaking, the assessment context should attempt to match the complexity of the target environment for the assessments to be meaningful.

An assessment is generated by collecting data about the agent, calculating a summary score about the agent's performance, and then comparing that score to established performance standards or benchmarks. For example, the Army Physical Fitness Test (APFT) assesses each soldier's muscular strength and cardiovascular endurance using three tests: push-ups, sit-ups, and a two-mile run. For each test, the soldier receives a score, which is rated on a 0–100 point scale. Individual test scores are computed by comparing the soldier's performance to age- and gender-specific standards. For example, to receive a minimum acceptable score (60%) on the push-up test, an 18-year-old male would need to perform 42 push-ups in two minutes; to receive the maximum score, that same soldier would need to perform 71 push-ups, again in two minutes. Similar scoring methods are used for the sit-ups and two-mile run. Since a minimum score of 60 is required to pass each test, each soldier's total PFT score can range from 180–300.

The assessment data can come from a variety of sources, such as self-reports, expert observer ratings, tools and technologies (e.g., simulators, radio networks), and sensors in the environment. Summary scores are computed using logical or mathematical formulae that codes, combines, transforms, and synchronizes the raw data to create a summary about the agent's performance (Freeman, Stacy & Olivares, 2009). Scores may be quantitative (e.g., number of push-ups, percentage of correct responses on a test) or categorical (e.g., a subjective "readiness" rating). The summary scores are then used to make some kind of decision (an assessment) about the agent. For example, a soldier who performs poorly on the APFT would likely be assigned remedial physical training.

Assessment Process Framework

The Army relies heavily on the results of human-performance assessments to ensure force readiness (e.g., Hawley, 2007). The Department of Defense (DOD) has made evaluation standards readily available online (e.g., Department of the Army (2004) Training and Doctrine Command [TRADOC] Pamphlet 350-70-4; DOD [2011] MIL-STD-46855A) and there are published handbooks (e.g., Boldovici, Bessemer & Bolton, 2001; Charlton & O'Brien, 2001) that describe procedures for conducting training evaluation and human factors testing to include the design, development, testing, reporting, and/or reuse of assessments. However, they are primarily high-level guidelines, with little or no specific details. Instead, individual assessment teams must make specific decisions and develop details about how to conduct assessments. Those assessment teams may vary greatly in terms of their experience and expertise in conducting assessment activities. Further, there is a plethora of strategies, techniques, tools, and instruments that can be applied. Even experienced assessors may differ significantly in terms of their methods. This leads to a great deal of variability in methods, impacting the both the quality and comparability of assessments. We offer the opinion that there is a clear need in assessment community for a guiding framework that describes the major elements of assessment related activities that are particularly relevant to the Army. Generally speaking, such a framework would serve to as a useful tool for both novice and expert assessors, providing high-level guidance for critical stages of the assessment process.

Figure 1 presents the Assessment Process Framework (APF), a conceptual framework that describes how assessment is accomplished across multiple phases: planning, execution (which includes data capturing, training aids, and monitoring), postprocessing data analytics, and cyclical improvement based on the observed results. Recognizing the futility of trying to develop a prescriptive "one-size-fits-all" approach, a high-level set of diagnostic questions is included in the framework to assist researchers and engineers – many of whom may have no formal experience in human-performance assessment – in navigating this complex and multifaceted process. At each stage of the process, users are presented with a series of questions or issues for consideration. Collectively, the user's responses to these questions guide them toward the best measurement approach that balances their project-specific needs with opportunities for reuse.

The APF was developed to be broadly applied across evaluation of technologies, studies of training interventions, the introduction of new work procedures and policies, the adoption of new organizational structures, and so forth. The process described here is consistent with best practices promulgated by the DOD (e.g., Bjorkman, 2008; Department of the Army, 2012) and NATO (DOD, 2002). Like that guidance, the APF process assumes that the assessment team has the expertise necessary to do the job, either themselves or by reaching out to credible experts.

As illustrated in Figure 1, an assessment typically begins with a specific task, usually from the sponsor or governing body. The assessment team must first ensure that they fully understand the objectives, constraints, and context of the task that has been presented to them. The team then plans the assessment. This involves developing the goals, approaches (or framework) and specific methods. The team then exercises these plans via a pilot test, simulation, or thought experiment. The formalized plans are then provided to the sponsor or governing bodies for review.

Upon receiving formal approval to proceed, the assessment team then develops assessment-related materials. This may entail designing or buying measurement instruments or stimuli and training the assessment team members who will act as field observers or data analysts. A pilot test is recommended as it may lead to revised and improved methods and materials. Next, the team executes the assessment, analyzes the assessment data, and issues a report of findings and recommendations to the sponsor, who in turn may make a decision or take action that is proportionate to the cost of the assessment. The decision may shape future assessment tasking from that sponsor. Feedback from the sponsors and from assessment team members supports an After Action Review (AAR), in which the assessment team identifies lessons learned and revises its assessment procedures to improve their future work. In the following paragraphs, we break the planning part of the assessment process down into its component parts.

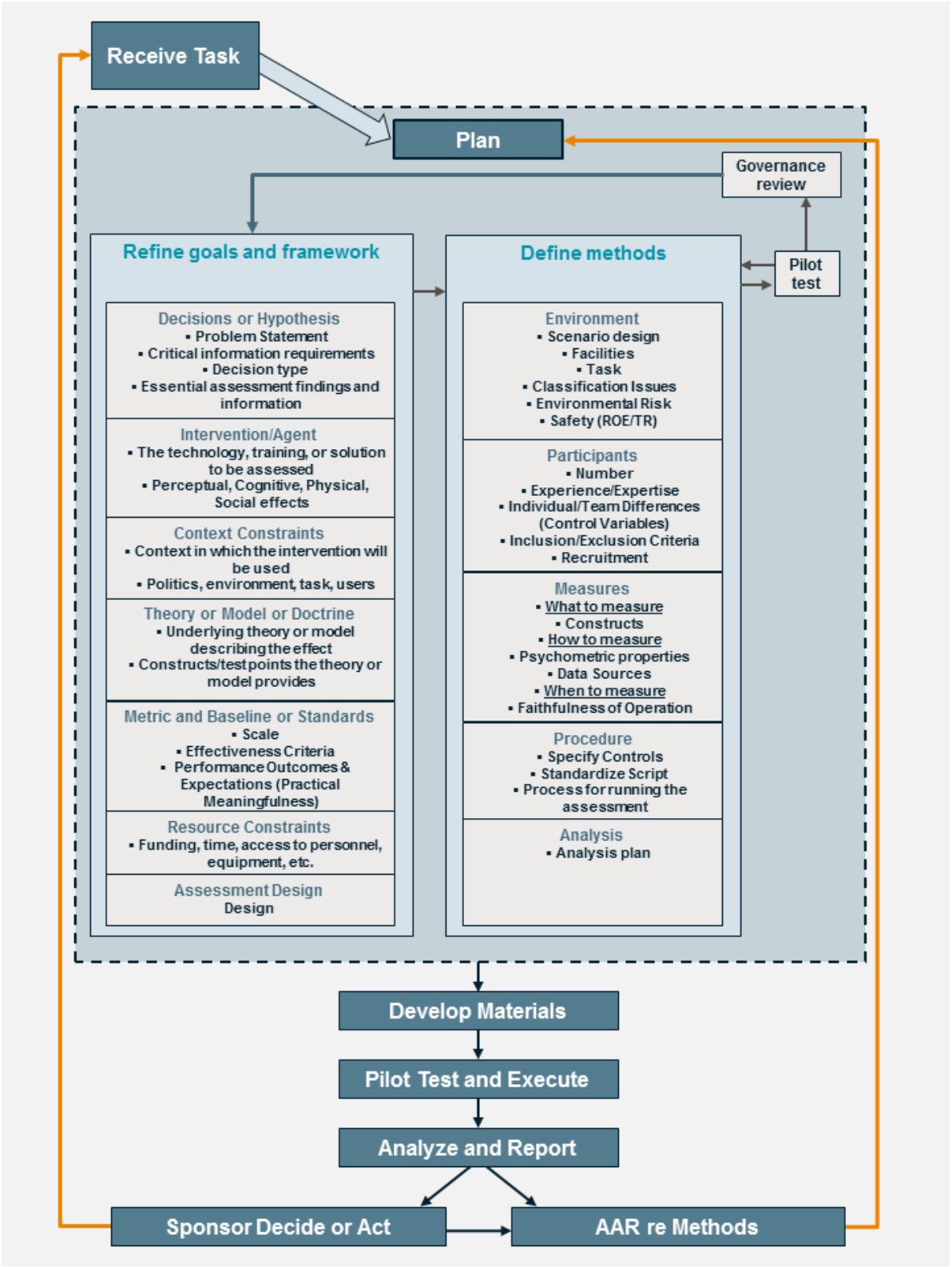


Figure 1. The APF.

Refine Goals and Framework

Upon receipt and clarification of an assessment task, the assessment team should refine its goals and develop its approach (or assessment framework). In this activity, the team may (in any convenient order), codify the decisions to be made by sponsors, the hypotheses to be tested, or the research questions to be explored; study the technology (or other augmentation, such as training or a new procedure) that is intended to enhance human performance; and consider the context in which it will be used.

The team should examine or propose a theoretical model that explains or predicts the effects of the performance augmentation; define the type(s) of information that an assessment must produce; and identify the baseline, standards, or metrics on which to judge (or, literally, to assess) that information. In addition, the team should consider any constraints on resources (e.g., money, time, facilities, and expertise) that may narrow the space of potential assessment designs.

Define Methods

Having established a useful framework, the assessment team then defines its methods. This entails discussing and making decisions about the environment in which users can apply technology (or other augmentations) to execute the tasks of interest, the participant sample that best represents the user population (given resource constraints), the measures that will be applied to the participants during task execution, the instruments that capture data to compute those measures, and the protocol for executing the assessment. In addition, the assessment team needs to consider how exactly the captured data will be transformed into measures and assessments, and how those assessments will inform recommendations. The product of this activity is the methods section in an assessment report.

It is advisable to pilot the assessment methods in some manner during the planning process. This is a rapid and inexpensive sub-step that identifies errors and inconsistencies in the methods. It can be executed as a thought experiment, an exercise over a conference table, a simulation, or a trial run in the operational environment. Following any refinement of the assessment method, the team should document its goals, framework, and methods, and provide these for governance review (in the upper right of Figure 1). With this documentation, sponsors can ensure conformance with their objectives and resources, both of which may change during or as a result of planning. In addition, assessments involving human research participants often require review by a Federally approved Institutional Review Board (IRB), which ensures that the protocols balance costs and benefits of using human subjects in research.

Example Use Case

In this section, we illustrate the application of the APF to a hypothetical, operationally motivated use case. Our intent is to demonstrate the types of diagnostic issues that an assessment team might identify and address at each step of the process. This presentation assumes a knowledgeable assessment team with expertise in areas such experimental design, statistics, and social science. The use case described here is relevant to Army programs of record and presents the assessment team with a diverse set of issues.

The use case involves an application based on a simulation to Mission Command Interoperability (SIMCI) automated course of action (COA) concept (Smith, Sprinkle, Powers, Xu & Knapp, 2015). The application is a planning tool that enables COA comparison within the Military Decision Making Process (MDMP). Currently, staffs make COA decisions based on mental and discussion-based COA comparison meetings. The proposed application in this use case would allow for Army staffs to easily draw out their COA in a planning tool and make decisions based on quantitative metrics enabled through constructive simulation. The assessment question becomes, whether that intervention, the COA comparison application, enhances staff performance in a meaningful way.

Receipt of Tasking

Upon receipt of tasking from the sponsor, the assessment team can clarify the sponsor's goals. The Heilmeier Catechism (Defense Advanced Research Projects Agency, 2016) is a useful list for this purpose: *What are you trying to do? How is it done today? Who cares? What difference will it make if you are successful*? Specific responses for this example use case might be the purpose of determining whether the COA automation tool improves the quality of mission planning vis-à-vis current methods. The Army organizations who care about the answers to this study include the TRADOC Capability Manager for Mission Command (TCM Mission Command) and the Program Executive Office Command Control Communications-Tactical (PEO C3T), because the proposed tool has the potential to improve mission plans, thereby saving lives, time, and materiel.

Refine Goals and Framework

After confirming the tasking that was provided to them, the assessment team needs to refine its goals and assessment framework by answering such questions as *What are the critical decisions, comparisons, or hypotheses to be tested? What are the likely contextual constraints? What relevant theories, models, or doctrine help inform the problem space? What is the critical metric or baseline for comparison?* Specific responses for this example use case might involve the critical decision of whether or not the Army should fund the COA automation prototype to bring it from stage 6.3 (Advanced Technology Development) to stage 6.4 (Demonstration and Validation). The project constraints might be that only five battle staffs are available to participate in the experiment, and that the experiment must be completed within the next six months. Critical theories that inform the problem space include models of teamwork (Kozlowski, Gully, Salas & Cannon-Bowers, 1996), trust in automation (Lee & See, 2004), and automation-induced complacency (Parasuraman, Molloy & Singh, 1993). Finally, according to the project sponsor, to advance the effort from 6.3 to 6.4, the COA automation tool results needs to improve COA quality by 25% and reduce COA development time by 25% versus current MDMP standards.

Define Methods

Next, the assessment team needs to operationalize its approach by responding to such questions as *What scenarios will elicit the critical behaviors or phenomena that are to be assessed? What assessment environment or facilities can support those scenarios? Do the participants require any special backgrounds, expertise, or access for the assessments to be meaningful? What measures address the constructs of interest? What instruments or sensors can capture data from those sources?* Specific responses for this example use case might include the scenario of a battalion-level armored engagement modeled on current National Training Center (NTC) pre-deployment exercises. The environment may involve one with resources similar to those available at the Mission Command Battle Laboratory (MCBL) at Ft. Leavenworth, KS. For the assessment to be meaningful, the participants need to include an intact battalion battle staff that has worked together for at least one full combat deployment. The measures include expert Observer/Controller ratings of COA quality, time to completion, and number of missed opportunities. The instruments required to perform the assessments include expert observer ratings, participant self-reports, and system-based measures (e.g., time to completion) from the COA automation tool.

Taken together, the questions addressed at these three stages should well position the assessment team to develop its assessment materials, scenarios, and data collection protocols; pilot test their methods prior to formal data collection; conduct their experiment; collect, analyze, and summarize the experimental results, and then present these results to the project sponsor or governance board for feedback and/or subsequent tasking.

APF and GIFT

As mentioned, the APF was developed as a broad guide for assessment across many domains and types. This section discusses the usefulness of the APF within the context of GIFT.

GIFT is a framework for assessment-driven instruction and instructional research. Accordingly, assessment is central to GIFT. Assessment-driven means that its three major functions involve assessment, and most of its software modules either generate or incorporate assessments. The major functions of GIFT include 1) authoring of instructional systems, 2) delivering instruction, and 3) analyzing those systems (Sottilare, Brawner, Goldberg, Holden & Smith, 2012). Each function entails assessment. Authoring engages instructional designers to design standards by which assessments are made and generate content that exercises the knowledge and skills to be assessed. Instructions assess and predict the state of learners and adapt training content and instructional strategy to their needs. In GIFT, learner state is broadly defined to encompass affect, cognition, performance, demographics, and learning history, each of which is a function of assessment. Finally, the analysis function produces assessments of the formative and summative effects of technologies, learning processes, and learning outcomes.

Four of the five GIFT software modules (or functional elements) perform or incorporate assessments to implement the GIFT functions (Sottilare et al., 2012; Sottilare, Brawner, Sinatra & Johnston, 2017). The sensor module transforms raw sensor feeds to processed data concerning learner affect and cognitive states for the learner module. The learner module assesses and predicts learner states as a function of 1) input from sensor module, 2) current performance data from domain module, and 3) demographics and historic performance data from an learning management system (LMS). The pedagogical module takes assessments from the learner module and generates new requests for assessment (e.g., to discriminate between hypotheses concerning learner state or effects of training treatments), as well as requests for feedback, instructional content, and instructional strategies. Finally, the domain module represents assessment standards, which it compares to live performance data to generate the assessments requested by the pedagogical module.

Example Use Case

The planning process defined in Figure 1 can inform the design of tutors that use GIFT. Assume that an eminent engineer, Emily, plans a design of two tutoring systems to learn which one accelerates the acquisition of technical maintenance skills including fault diagnosis, repair, and testing.

During planning, Emily *refines the goals and framework* of her instructional project. At this planning stage, she may take the following actions:

- Identify competing *theories* of learning, one that correlates learning with depth of processing, another with time on task.

- Define *hypotheses* that reflect those theories and that drive tutor design. For example, one tutor might incite deep processing by leading students to induce mental models of devices they must repair. Another tutor might present diagnostic rules to students and engage them primarily in memorizing and applying those given rules.

- Select *metrics* of performance on diagnostic, repair, or testing tasks from current training requirements or operational requirements.

- Consider the *context* in which the tutors may be used. Perhaps she concludes that technical specialists should use the tutors while on military deployments (e.g., in the ship's mess or a base library in theater).

The goals and framework that Emily has identified in the previous step influence the methods she defines. She methodically plans how she will do the following:

- Design a maintenance task *environment* that is relevant to personnel from all of the services, perhaps one that focuses on common machinery such as electrical motors.

- Characterize her *participants* in ways that enable her to distribute her tutors systematically and test their effects on these different groups. For example, she might partition students by expertise – as novice, journeyman, expert, or master (Charness, Feltovich, Hoffman, Ericsson, 2006) – to test for differential effects of each tutor given prior knowledge.

- Define *measures* of the frequency and timing of use of her tutors, the overall effects of each tutor on expertise, and the contribution of each exercise to that effect.

- Define technical and logistical *procedures* for capturing data from field units.

- Define *analyses* of the main effects of each treatment and the interactions between the treatments and initial expertise.

This example, we hope, illustrates the benefits that an instructional researcher or designer gains by systematically planning assessments using the framework presented here.

Recommendations and Future Research

The UMMPIREE research program is in its early stages of identifying methods and tools to enable systematic assessment of training and augmentation interventions for the Army. The APF was a first step in identifying the critical steps in conducting assessment for a variety of purposes including assessment with the context of technology-enabled learning environments. This framework can serve as a high-level guide to those conducting military-based assessments and is a step in the right direction. However, we have identified additional challenges to achieving the goal of systematic assessment across the Army. First, a standardized method of assessment must be relatively flexible to accommodate the range of assessment challenges. In particular, manual techniques will be appropriate for some tasks (such as developing theory and measures), technologies will fit others (such as automated capture of experimental data), and both must include options for many (e.g., pilot tests may be conducted with automated simulations or live enactments). Second, to the extent that the vocabulary of human performance assessment can be standardized, human-readable and computer-executable languages can be developed to automate assessment tasks. Third, selection of the "best-fit" measures can be furthered if measures are well defined and catalogued for discovery and reuse by different assessment teams. Finally, the goal of increasing confidence in experimental findings is, itself, a measurable goal. Accordingly, a process and authority must be established to monitor and refine use of the framework we have presented.

References

Bjorkman, Eileen. (June 23, 2008). Joint test and evaluation methodology (JTEM) overview. Presentation made at a meeting of the NDIA Systems Engineering Division Developmental Test and Evaluation (DT&E) Committee. Accessed 7/5/16 at http://www.ndia.org/Divisions/Divisions/SystemsEngineering/Documents/Committees/Developmental%20Test%20and%20Evaluation%20Committee/2008/June%20Committee%20Meeting/6-24-08%20NDIA%20DTE%20Committee%20Charts.pdf.

Boldovici, J. A., Bessemer, D. W. & Bolton, A. E. (2001). The elements of training evaluation. Army Research Institute for the Behavioral and Social Sciences, Alexandria VA.

Charlton, S. G. & O'Brien, T. G. (Eds.). (2001). Handbook of human factors testing and evaluation. CRC Press.
Charness, Neil; Feltovich, Paul J.; Hoffman, Robert R.; Ericsson, K. Anders, eds. (2006). The Cambridge Handbook of Expertise and Expert Performance. Cambridge: Cambridge University Press. ISBN 9780521840972.
Defense Advanced Research Projects Agency (DARPA). (2016). The Heilmeier Catechism. Retrieved from http://www.darpa.mil/work-with-us/heilmeier-catechism.
Department of Defense (2011). Department of Defense Standard Practice: Human engineering requirements for military systems, equipment, and facilities (MIL-STD-46855A).
Department of the Army. (2012). Army Regulation 5–5: Army Studies and Analyses. Washington, DC, Department of the Army. Retrieved from http://www.apd.army.mil/pdffiles/r5_5.pdf.
Department of the Army (2006). Test and Evaluation Policy (Army Regulation 73-1). Washington, DC: Department of the Army Headquarters.
Department of the Army (2004). Systems Approach to Training: Evaluation (TRADOC Pamphlet 350-70-4). FT Monroe, VA: Training and Doctrine Command.
Department of Defense. (2002). NATO Code of Best Practice for C2 Assessment. ISBN 1-893723-09-7
Freeman, J., Stacy, W., Olivares, O., In Cohn, J., Schmorrow, D. & Nicholson, D. (2009). Assessment for Learning and Development in Virtual Environments. The PSI Handbook of Virtual Environments for Training and Education: Developments for the Military and Beyond, 236–250.
Hawley, J. K. (2007). Looking Back at 20 Years of MANPRINT on Patriot: Observations and Lessons (No. ARL-SR-0158). Army Research Lab, Adelphi MD.
Kozlowski, S., Gully, S., Salas, E. & Cannon-Bowers, J. (1996). Team leadership and development: Theory, principles, and guidelines for training leaders and teams. Advances in Interdisciplinary Studies, 3, 253–291.
Lee, J. & See, K. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46, 50-80.
Parasuraman, R., Molloy, R. & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." International Journal of Aviation Psychology, 3, 1-23.Smith, M., Sprinkle, R., Powers, J., Xu, J. & Knapp, M. (2015). "Fixing" the Military Decision Making Process (MDMP). In Proceedings of the 2015 Interservice/Industry Training, Simulation, and Education Committee (I/ITSEC). Paper No. 15220. Arlington, VA: National Training and Simulation Association.
Sottilare, R.A.; Brawner, K.W.; Goldberg, B.S.; Holden, H.K.; Smith, P.R. (October, 2012). *The Generalized Intelligent Framework for Tutoring (GIFT).* Retrieved from https://gifttutoring.org/attachments/download/152/GIFTDescription_0.pdf.
Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017. DOI: 10.13140/RG.2.2.12941.54244.

CHAPTER 20 – Assessment in Intelligent Tutoring Systems in Traditional, Mixed Mode, and Online Courses

Anne M. Sinatra[1], Scott Ososky[1,2], and Robert Sottilare[1]
US Army Research Laboratory[1]; Oak Ridge Associated Universities[2]

Introduction

As technology has continued to develop, there have been new and creative ways to use it in the classroom. Classes have transformed into learning environments that rely not only on the in-person instructor, but have many other resources easily available for students. At the college level, courses can vary between different modes of instruction: traditional in-person, mixed mode (classes with both an in-person and online component), and online. Students routinely bring laptops and tablets to class to both take notes and work on their assignments. Further, there are often computer labs available for students to use for assignments at universities. In some cases, these labs can be reserved for specific classes and provide all students with the opportunity to engage with a computer during class time.

Even traditional in-person courses often have online materials available for students through learning management systems (LMSs) such as Blackboard or Webcourses. Students often engage in message board discussions with their classmates and download resources such as class PowerPoints from the sites. Grades are often times managed digitally and are available to students through logging into the system. Many LMSs also have built in assignment submission sections that in some cases can check the student submitted work against others for plagiarism. Many of these resources are used by instructors as they design and supplement their courses. These resources can be combined with in-class instruction to build a learning environment that provides students a self-regulated method of engaging with additional educational materials. In the case of mixed-mode or online course, the students heavily rely on these LMSs and primarily receive their instruction through engaging with the computer-based materials. In these types of classes, the instructor's goals when developing materials will be different than for in-person classes, such that the online materials will need to be deeper and more self-explanatory. In online and mixed-mode courses, the instructor moves from a lecturing capacity to a facilitator and subject-matter expert who engages with students when they turn in assignments, when they seek help and when they have questions. The online class environment is much more self-directed and requires the students to engage in self-regulated learning and be more aware of what knowledge that they have (metacognition), which does not necessarily come naturally to all students (Schraw, 1998).

Discussion

Among the different methods that can encourage the engagement of students and further provide the student with tailored learning opportunities are intelligent tutoring systems (ITSs). ITSs have been found to be as effective as a live human tutor working directly with a student (VanLehn, 2011). In many cases, students do not seek out in-person tutoring or fail to even engage with instructors during office hours to improve their performance. ITSs provide a means of offering personalized tutoring to an individual that is tied directly to the student. Tutoring can occur based on individual difference characteristics, previous experiences or knowledge, current performance, among other variables that are at the discretion of the author of the ITS. ITSs can be used in conjunction with a formal course, as a required assignment, or as a course component. ITSs can be powerful tools for educators, as they can target in on specific material or components of material that students are having difficulties with. Some ITSs can engage students in a reflective conversation or try to correct misconceptions that they have. The progress that the student makes on the ITS

materials can also be recorded and used to make further decisions about materials that the student is to receive.

Courses which include computer-based components may have students engage with premade materials that linearly walk them through the learning process or provide practice. ITSs provide a means to give students personalized learning without needing to have a human tutor involved in the situation. There is upfront work that requires planning and creation of course materials that will send individual students down different remediation paths. Therefore, time is spent on planning and coming up with alternative ways of teaching the material to be learned. Similarly to the development of a fully online class, a great deal of time is spent up front preparing the materials that will be used in an ITS. However, after that initial time, the course does not need to be redeveloped and can just be monitored and changed as needed.

Currently, there are a few major hurdles to ITS use in the classroom and for assessment. The first hurdle is that creating an ITS is time consuming and often requires specialized skills such as computer programming knowledge. Secondly, ITSs are not always easily accessible from the internet and are not always available in mobile platforms. Thirdly, ITSs are costly to produce and tightly coupled to the material that they are teaching (Sottilare, Goldberg, Brawner & Holden, 2012). Projects such as the Generalized Intelligent Framework for Tutoring (GIFT) are making an effort to tackle these issues. GIFT is an open-source, domain independent ITS framework (Sottilare & Holden, 2013; Sottilare, Brawner, Sinatra & Johnston, 2017). GIFT is now available online (at https://www.gifttutoring.org) and can be accessed by both teachers and students over the internet. Further, GIFT provides a set of web-based authoring tools that are domain independent and designed to be straightforward to use (Ososky, Brawner, Goldberg & Sottilare, 2016). Care has been put into the design of the tools to ensure that subject-matter experts and instructors can create their own adaptive tutoring systems without the need for a background in computer science. The generalized tools that have been created for GIFT allow for the reuse of the materials and the ability to quickly and efficiently create adaptive content. Additionally, GIFT is open source and freely available for use by instructors.

Assessment in ITSs

ITSs can be used to not only supplement in class lectures, but also can provide a means of assessing students. ITSs such as those made with GIFT can include multiple-choice questions, and other assessment methods that can be used to determine the progress of an individual student and the understanding that they have of the subject. Assessments in ITSs can either be at the end of overall lessons or they can be leveraged to provide questions during the learning process that can lead to individual remediation based on the performance of the student, while still retaining a record of the original scores. ITS authors can determine the point values and difficulty levels of questions to set up the types of remediation that they want their students to receive and how they want assessments to be graded.

There are many approaches that instructors can use to incorporate ITSs into their classes (Sinatra, 2015), for instance, completion of an ITS by a student can be used as checkmark completion grade or as a more fine-grained assessment that can count toward the individual student's semester grade. Another approach is to use it as a required remediation tool prior to an in-course exam, which could be compared to the performance of students that did not have that specific study tool to examine the effect of it. Thirdly, ITSs could be used as an assessment by having students design, create material for, and author their own ITSs using a framework such as GIFT. This practice could be beneficial to students specifically in the field of education, but also graduate students in general who may end up teaching their own classes. Even those who do not intend to teach a course could benefit from interacting with the authoring of an ITS, as it would get them to think about what types of questions could be asked about material. Students could be assigned a chapter and asked to think critically about what could be tested on in that chapter and create their own

ITS with questions to assess it. Students could also be asked to create their own experiments to determine how others would learn most efficiently, by having different methods of assessments and running participants through them. See Table 1 for examples of techniques in which ITSs could be used for assessment.

Table 1. ITS assessment techniques.

Student Population	Assessment Approach	Benefit
Graduate students	Ask students to create their own ITS to help them practice for teaching courses.	Real-world experience with creating courses and assessments
Undergraduate students	Assign a chapter to students and ask them to create components of, or an entire ITS to teach that chapter.	Encourages students to engage in metacognitive assessment of themselves, and reflect on material that they learned.
Graduate or undergraduate students	Create assessments in the ITS for use for in class grades.	Assess student knowledge of material.
Graduate or undergraduate students	Educational research	Teachers create alternate means of teaching and compare to other years/methods or students are asked to create studies that do so and bring in others to participate.

ITSs can be extremely useful to instructors. See Table 2 for examples of the benefits of ITSs in different modes of classes. There are a number of considerations that arise when implementing the use of an ITS in a classroom environment. There will be an impact to classroom management, to the amount and type of content needed, and the information that an instructor will want about the students during class time.

Table 2. Benefits of using ITSs in different types of classes.

In-Person	Mixed-Mode	Online
* Supplements in-person lectures. * Provides opportunity for clarification on points that were not completely understood. * Can be designed to encourage metacognition and allow self-assessment of class material. * Can provide an automatically graded quiz or exam to the instructor, which can then be examined to see what questions were answered correctly or incorrectly.	* Provides engaging opportunities to build on material provided in class. * Can be used to introduce new material before it is discussed in class. * Can be used to assess understanding of in-class material, and results can be gone over with students at the next class meeting. * Students can generate their own tutors for chapters that were presented in class, which can then be shared with each other in addition to the instructor.	* Increases engagement. * Assists with self-regulated learning. * Demonstrates personalization. * Students can receive customized feedback without needing to see an instructor in-person.

Considerations for Using an ITS as a Learning Tool in the Classroom

Although many ITSs have been designed to provide one-to-one tutoring experiences, they are being applied in traditional classrooms more often to augment instructors/teachers and support concepts like flipped classrooms where classroom lectures are viewed at home and typical homework elements (e.g., assignments or projects) are done in the classroom under the supervision of the instructor (Tucker, 2012). ITS capabilities exist today to allow teachers to deploy parallel assignments to their students to support independent, tailored learning experiences in the classroom where each student can proceed at their own pace. To manage the independent lessons of their students, ITS developers have designed dashboards to represent the students' progress toward objectives, their domain competency, emotional state, level of engagement, and other data analytics about the time each student spends using each piece of content.

While ITSs have value in the classroom in supporting self-paced individual exercises, they also have limitations that should be considered prior to their application. First, allowing students to progress at their own pace means that content must be developed/curated to meet all levels of performance. In particular, low-performing and high-performing learners have different content and feedback needs. Low performers may require multiple passes at content to reach a desired level of competency. High performers may consume content at such a rate that they run through the available content and are seeking more challenging material. Considerations should be given to allow high performers access to material in more challenging subsequent modules. Mechanisms should be available in the ITS to allow this type of assessment of learning trends.

A second consideration is in supporting the assessment of collaborative learning, an educational approach to instruction where groups of learners work together to make a decision, solve a problem, complete a task, or create a product (Bruffee, 1999). Considerations should be given to provide access to data to perform team assessments. The use of low-cost, unobtrusive sensors and mechanisms to capture learner actions on a computer, on a mobile device, or within a simulation will allow ITSs to capture physiological and behavioral data needed to classify individual and team states based on an evidence-based approach.

A third consideration is in managing the instructor workload associated with using ITSs in a flipped classroom. Inevitably, students may find themselves in situation where the ITS cannot support their learning needs (e.g., unable to answer a question or unable to provide new content). At this point, it may be necessary for a live instructor to intervene. With 20–30 students, how might a single instructor prioritize these needs

for interaction? The answer is likely to be some type of computer-based dashboard where the status of each student is tracked and alerts inform the live instructor as to the type of issue the student is encountering.

Current effort is being made in GIFT to create an instructor dashboard to assist with the identified issues, and to help the instructor monitor the progress of students. The design of the dashboard is expected to be customizable and domain-independent.

GIFT Instructor Dashboard

The previous sections have discussed ITSs and how they have been used for assessment in classroom settings and how interacting with ITS authoring tools can teach students about tutor development. This section focuses on an instructor's needs supporting *assessment* with an ITS, specifically the application of an *instructor dashboard* within the ITS platform. GIFT is used as the exemplar ITS platform for the purposes of this discussion.

Instructor-Centered Design

An instructor dashboard for GIFT, and ITSs in general, should be able to serve an instructor's goals. Instructors may have one or many goals related to the course, the students, and their own planning. Goals of interest to an instructor may include increasing student learning and reducing attrition, enabling effective time management given an instructor's various responsibilities, designing effective materials and courses, and enabling effective information management for external reporting (Siemens et al., 2011). Those goals are not mutually exclusive and those goals may change over time. Instructor dashboards do not directly meet those types of high-level goals; rather, they enable the completion of individual tasks that ultimately serve those goals.

Let's consider, for a moment, what tasks may be accomplished by an instructor dashboard. An instructor may want to track the progress and performance of students over the duration of a course's administration, including predictions of final grades. The instructor may similarly want to track learner behaviors and attitudes regarding the course over the same period of time. There may be interactions between students or between students and the instructor that need to be monitored within the dashboard. That may include providing feedback to students, as needed (Holman, Aguilar & Fishman, 2013). Taking a broader-picture view, the instructor may want to monitor student performance within the context of a learner's education career or compare performance against some institutional set of standards. Also, the instructor may want to evaluate the design of their course (Grover, Pea & Cooper, 2014).

Furthermore, the learning content, assessment material, and *adaptive* interventions that populate a tutor within an ITS may greatly differ between modules, even within the capabilities present in GIFT. Incorporating adaptive information into the instructor dashboard is a challenge not found in traditional computer-based instruction. The instructor will likely want to know what adaptive paths a learner encountered. That might include the items that were dynamically presented from a question bank, as well as the difficulty of those questions and the course concepts addressed by those items. Adaptations may take the form of real-time assessment, such as those found within a practice environment, like an educational simulation environment. Adaptation in GIFT may also be found within a student's conversation with a synthetic agent, or the dynamic selection and presentation of content to the learner, which occurs in discrete time. Given the variety of potential goals the instructor may have in conducting assessment-related activities and the ways in which a GIFT tutor can be adaptive, it would seem appropriate that an adaptive tutoring platform has the capability to configure adaptive instructor dashboards. Those dashboards will need to react to the data sources available to the system and output the information in such a way that the instructor will be able to conduct the relevant analyses in order to address their needs.

Data Sources

Data are at the heart of the instructor dashboard. Under ideal circumstances, GIFT will integrate data from both external and internal sources. External sources might include information from individual learner models, such as previous grades from related and/or prerequisite courses, learning preferences, personal interests, or other relevant academic information (Sabin, 2012). Some of this information could come from a centralized system (e.g., Blackboard, edX) using a standardized format (e.g., experience application programming interface [xAPI]). The dashboard should also account for instances where this information may need to be manually integrated (or perhaps by means of a questionnaire presented to the learner at the start of a module or course). In any event, prior information may be inconsistent between students, incomplete within a student, or simply not available at all. A learner dashboard in GIFT should be flexible enough to work with external information when it is available, but still draw meaningful conclusions for the instructor when external information is not available or not complete enough to be useful.

Internal data sources to populate a GIFT learner dashboard for assessment can come from many sources within GIFT. Learner interactions with surveys will naturally produce data suitable for assessment within a learner dashboard. GIFT, however, provides additional information with respect to survey interactions. Survey questions can have partial scores or multiple correct answers. Data may also include metadata about the difficulty of the question and the concept(s) being assessed. If a survey was generated from a question bank, the order and selection of questions presented to the learner may also be of interest to the instructor. GIFT also includes options for non-scored (e.g., demographic) surveys, which may add additional ways to slice assessment data within a dashboard.

GIFT tutors are made up of individual parts, known as *course objects*. Course objects represent presentations of interactive and static content, or contain the logic for advanced interactions including external applications. For static content such as text, images, or video, data might be collected about the time spent looking at these materials (dwell time). GIFT's *structured review* course object allows learners to review a previous assessment; information regarding the pages that were accessed, and time spent reviewing the quiz could be recorded for further analysis. In addition to the dynamic display of survey materials, GIFT's *adaptive courseflow* course object can dynamically display media, documents, and web information based on characteristics of the content and learner, respectively; the instructor may want to know what material was presented, as well as how often remediation was given, to better understand survey/quiz scores. GIFT's *external application* course object leverages *real-time assessment* logic to exchange data with software such as simulators and games; this is a potentially rich data source that might include changes in learner performance states, learner behaviors within the external software, or strategies, tactics, and feedback invoked by GIFT during the interaction. Finally, GIFT also supports the use of sensors (e.g., physiological devices) which are most commonly used as part of a real-time assessment.

Data Analysis and Presentation

Given the types of assessment-related questions an instructor may be trying to answer and the various data sources available to them, determining the so-called *best* way to present relevant information to the instructor may be something of a moving target (Brown, Lovett, Bajzek & Burnette, 2006). Additionally, the level at which the information is presented may be just as important as the source data itself. An instructor may prefer a presentation of raw student data, perhaps when the class size is small or when it is required based on the task to be performed. Aggregate or descriptive statistics may be preferable with larger class sizes, or when analyzing trends within the class or comparing against other populations. Further, an instructor may want to view some type of predictive analytics, during the administration of a course, to more quickly identify students that may be struggling based on some classification criteria (Siemens et al., 2011).

Statistics, data analytics, and visualization are all relevant parts of the instructor dashboard; however, there is one important caveat when approaching these data. Specifically, all data points represent some aspect of an actual student engaged within a particular course. In practice, this means that anomalies in the data, such as outliers, unusual, or otherwise missing values cannot simply be discarded or excluded from analysis. An outlier may indicate a student that is performing either highly above or below the class average. Investigating such data points in greater detail may yield useful information that can be used to either identify aspects of the course that are working, a student that may need more assistance than what is provided within the course or even the potential to identify behaviors consistent with cheating. Unusual data might include high and low time spent in various activities. While these are not necessarily indicative of any specific behavior, it might be to the instructor's benefit to follow up with a student that may potentially be skipping through content or perhaps leaving the tutor open while walking away to do something else (such as eating). The latter is relatively harmless, but the former may require additional instructor intervention. Again, the purpose of the dashboard is to allow the instructor to quickly and efficiently get to the ground truth of how students are progressing through the course, and provide support to those students that may require additional interventions.

Conceptual Design of GIFT Instructor Dashboard for Assessment

Given the prior discussion, it is worth reiterating that adaptive tutors should be supported by adaptive instructor dashboards. This means that an instructor should be able to call upon the tools needed, but the system should also be semi-autonomous in preparing the layout of the dashboard and generating the appropriate visualizations. This section presents some conceptual notions of an instructor dashboard (Figure 1) and describes how an instructor may interact with the interface to help them accomplish tasks supporting their overall work goals.

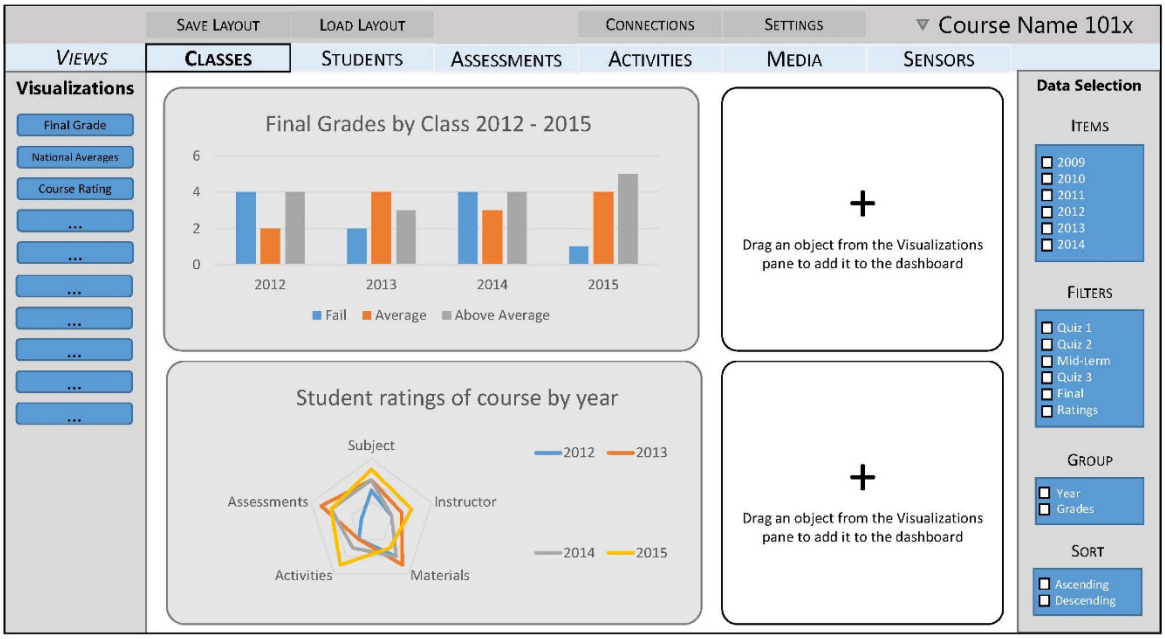


Figure 1. Conceptual sketch depicting modularity of dashboard divided by different elements associated with the course. The image above is intended to be more conceptual than it is prescriptive of a specific layout and/or user interface design.

An adaptive instructor dashboard can be enabled by a modular interface. This discussion starts from the center and works outward (both metaphorically and literally). The center/primary section of the interface is where the instructor would generate visualizations and interact with data. This area is intended to be entirely configurable to suit the instructor's tasks or questions to be answered with the data. However, the burden of building the visualizations should not be with the instructor. Instead, GIFT should provide a set of components that can be added to the dashboard (e.g., drag and drop) and preconfigured with a default set of options. Since the instructor dashboard resides in GIFT, it would seem practical that appropriate visualization modules would be made available to the instructor based on the known course object types in a specific tutor and the data sources that those course objects can provide. Providing preconfigured visualization modules will increase the efficiency of the tool, but at the potential expense of flexibility for power users. The ability to build one's own custom charts and graphs could be added as an advanced feature in a later version of the dashboard.

Once visualization modules have been added to the dashboard panel, the instructor would be able to rearrange and resize the panels to increase their readability and highlight their relative importance to the instructor's inquiries. Once that has been completed, the instructor is now ready to begin exploring the data. The "data selection" area has been modeled after a database query. Clicking on a visualization panel should populate the available options in the data selection area. From that area, the instructor should be able to specify what data they want to see, as well as how it should be displayed, grouped, and sorted.

Application of Instructor Dashboard

Even in regard specifically to assessment related goals, an instructor may have many different questions they need to answer: How did student A do? How did the class do in comparison to last semester's class? Are there items on the quizzes that students are consistently missing? Does my question bank need more or less material? Are the lesson materials supporting the instructional goals? To those ends, it would seem appropriate to introduce the concept of viewing data at the object level (Figure 1, 2nd level menu bar), which includes students and classes, but also may directly focus on specific assessments, media content, or practice environments (e.g., games or simulations). For example, imagine that an instructor analyzes the ratio of correct to incorrect responses for a particular quiz. The instructor finds that students are performing poorly only on a specific concept (Concept 1 in this scenario); the instructor decides to investigate how much time students spent on average, with the various lesson materials for the course concepts (Figure 2).

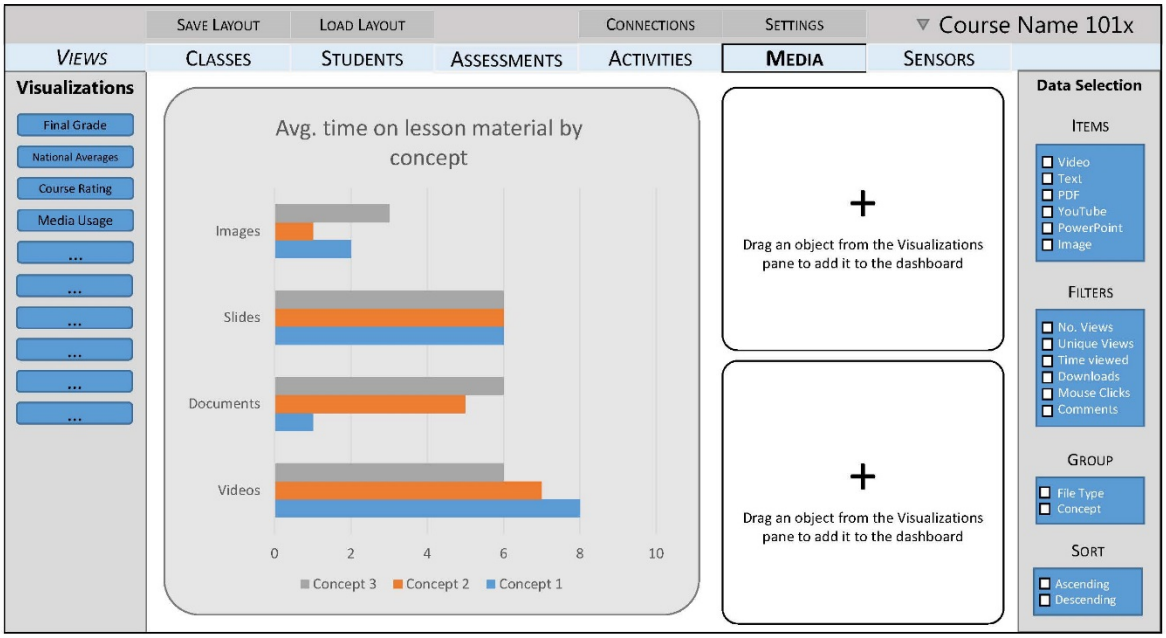


Figure 2. An object-oriented approach to high-level dashboard creation enables the ability to answer different types of assessment related questions within an instructor dashboard. Note that the "data selection" area changes based on the elements under investigation.

The instructor notes that students spend, on average, far less time with the documents (e.g., PDFs, text, web pages) associated with Concept 1 compared to the other lesson concepts. This might suggest that students spent less time reading the lesson material, or perhaps the tutor simply did not present that specific yet important material during the adaptive portions of the course. Either way, the instructor dashboard is intended to arm the instructor with the data and knowledge necessary to make better decisions about supporting students and improving the design of their assessments and courses.

Additional Interactions

There are a couple of other elements of this conceptual dashboard that can increase the efficiency with which instructors can accomplish their goals. First, the interface should allow the instructor to save the personalized dashboards that are built, presumably to revisit them at a later time or use them with other courses. Sharing dashboard layouts might also be on the short list for a second version of this tool.

The dashboard will also need user interfaces to guide instructors through the process of connecting to external data sources, such as an online LMS or an offline spreadsheet. Despite the technical complexities that this process may introduce on the system backend, the system should endeavor to provide a guided, semi-autonomous experience for integrating these data, even if the initial set of available connections is limited. Additional connection logic for other external data sources can be added at a later date. The interface should also offer the option to enable, disable, or disconnect these data sources in the event that data are no longer needed or wanted.

Finally, the instructor dashboard should eventually allow the instructor to navigate through the dashboard by making the visualizations interactive. For example, clicking on a specific rating attribute from the bottom chart of Figure 1 should allow the instructor to view the raw data, including any open-ended comments that

were made regarding that dimension of the survey. Similarly, clicking on one of the media types in Figure 2 should present the instructor with a list of the specific files, URLs, etc., that makes up that category of media, for the specific concept clicked-on. In addition to moving vertically through the data, it should eventually be possible to move horizontally through data or *pivot* between the different element types at the top-level navigation. For instance, an instructor may wish to quickly pivot from media types to concept-related information by clicking the relevant elements of the chart in Figure 2. doing so would change the data selection options and toggle the top level navigation automatically into a different element type.

Summary

Adaptive tutors require adaptive instructor dashboards to accommodate the varied data sources that may be found within a tutor. To effectively analyze assessments within a course, it may be necessary to consider other elements of the course, such as supporting materials and student behaviors. A conceptual model was proposed that included a semi-automated modular dashboard, leveraging preconfigured visualizations and dynamically updating data selection tools. It was recommended that a robust yet limited set of efficient options should initially be prioritized over advanced, granular flexibility. The instructor dashboard should also provide the ability to connect with external data sources, save/load/share dashboard layouts, and eventually provide a visual navigation capability by interacting directly with the visualization panels in the dashboard.

Recommendations and Future Research

General Recommendations for GIFT

In the current form, GIFT and other ITSs can be used in classroom environments for assessments. In the case of GIFT, it provides a means of recording all student performance/actions in the GIFT system and saving it to external log files that can have data extracted from it at a later time.

As GIFT continues to develop, there are a number of features that will be helpful to instructors that could be considered:

- Breaking down the questions that students got right or wrong into a report
- Statistics for the percentages of students that got questions right or wrong
- Ways of allowing students to provide feedback about the material
- Sending student grades directly to the teacher
- Exporting data directly to LMS formats such as Webcourses

The development of an instructor dashboard for GIFT will begin to address some of the above features. In GIFT's current form (pre-user roles), students can login anonymously and engage with the course or they can import a course into GIFT. The implementation of user roles and permissions is the next step, so that students in specific courses can see a course appear in "My courses" without needing to specifically import it. There should also be a mechanism that would send the student scores to a teacher's gradebook in the LMS. GIFT is in a stage where the concept of user roles and permissions is still being developed. Ultimately, the students and instructor login interfaces will need to be different. The instructor dashboard that is currently being developed will be a great resource for an instructor who wishes to use GIFT for assessment. Additionally, the functionality of the dashboard and information provided on it makes it useful for in-person, mixed mode, and online course instructors. With regard to a student interface, ideally, students should be limited to the courses that they are enrolled in and then the information should be automatically sent to the appropriate instructor, instead of just generating log files on the local computer or the cloud.

Unless they have specifically created a course, the student should not have the option to edit a course. Teachers should have options in regard to visualizing the grades that students received. Their tools should allow for reviewing test answers/percentages correct, and helping them to make decisions about good or bad questions. In the current state, GIFT is primarily populated from already generated log files. However, if a teacher determines that there is a problem with a specific question, it might be beneficial to be able to make an adjustment after participation has occurred, where the grading of the question is updated in the system. If this type of mechanism is not implemented, then it is important to provide an instructor's GIFT handbook so that they know the ins and outs of what GIFT does and so that their expectations are in line with GIFT's capabilities.

GIFT currently has a method for taking questions from a user generated bank and compiling them into tests that cover concepts. While the questions that were asked are in the log files, a way to visualize these and provide a record of the asked questions will be needed for instructors. Instructors want to know what questions students are doing well on and make sure there is equal difficulty in the assessments that everyone is receiving. Therefore, it is important than an export feature be implemented that shows the student's progress, the grade they got on the question, and other information that will assist them in their assessments. GIFT's instructor dashboard will be a step forward to assist instructors in visualizing student progress and activities.

Recommendations for Instructor Dashboards

To effectively analyze assessments associated with classroom-based adaptive tutors, adaptive instructor dashboards are required. GIFT already generates tutor interaction information that can currently be found within an event reporting tool; however, its usability is currently limited to researchers and power users. An instructor dashboard should seek to automatically organize this data for the benefit of the instructor. Preconfigured visualization options should be provided, prioritizing depth and usability over flexibility.

Additionally, a number of hypothetical instructor questions were described in the instructor dashboard section. These questions generate tasks that an instructor will perform within the dashboard to answer questions that ultimately serve broader goals (e.g., effective lesson planning, meeting organizational objectives). At this time, it is not clear that GIFT natively generates appropriate data to power all of the proposed aspects of the dashboard. For instance, GIFT can provide some monitoring of student interactions with PowerPoint files, but this is inconsistent with the data generated from viewing a media course object. Surveys generate data regarding student performance, but not necessarily the amount of time spent on the survey, page, or question. Conversely, physiological sensors generate vast amounts of data, often requiring a customized viewer for each, which may be difficult to translate into the instructor dashboard interface in a meaningful way. It is recommended that instructor requirements guide the design of the dashboard and, in turn, inform the type of data that should be generated when a student interacts with various course objects in a GIFT tutor.

Instructors would benefit from a function in GIFT that generates student data through simulation. This is relevant not only to quality assurance in course design, but also the instructor's planning for a dashboard. With simulated data, the instructor could activate and explore different visualization modules in the dashboard to ensure that the questions they will later pose with actual students can be answered by the data provided. An instructor may similarly use a dashboard layout that was used for a different course or one that was shared with them in evaluating its usefulness for the current course. Simulated data can also be used to test a new dashboard with an external connection to ensure that the imported data will add value to the data generated internally.

For ITSs to be more useful in the classroom, first, we recommend additional investigations into automated authoring methods to reduce the ITS developer burden and more easily expand the type and quality of content available to different levels of performers. A second recommendation for ITSs in the classroom centers on the development of low-cost, unobtrusive sensor suites that include the sensor hardware and associated classification models, which use individual learner behaviors and affect along with team behavioral markers to identify individual learner and teamwork states like team cohesion or collective efficacy. Finally, a third recommendation related to using ITSs in the classroom is the development of an instructor dashboard, which could be used for multiple task domains, but is primarily tied to cognitive tasks and assessment of cognitive task performance. This dashboard should also include a visualization of the long term learner model attributes which contribute to success in the current domain under tutoring.

Conclusions

ITSs provide many benefits to instructors who wish to use them for assessment. While the execution of using ITSs in the classroom will vary based on the level of the student (e.g., high school, college, etc.) and the mode of the class (in-person, mixed mode, online), there are many useful generalizable features of ITSs that instructors can use. The inclusion of an ITS in a classroom environment will lead to adjustments that need to be made by the instructor such as managing the classroom, authoring additional material, and determining how the ITS will be incorporated for grades. Using an ITS framework such as GIFT provides instructors the flexibility to use GIFT either in-person or online, and to not only create, but reuse parts of their ITSs for different classes. The addition of user rules, and the development of an instructor dashboard will make GIFT an even more powerful tool for instructors to use in or out of the classroom.

Acknowledgements

The research described herein has been sponsored by the US Army Research Laboratory. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Brown, W. E., Lovett, M., Bajzek, D. M. & Burnette, J. M. (2006). Improving the feedback cycle to improve learning in introductory biology: Using the Digital Dashboard. *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 1030–1035.

Bruffee, K. A. (1999). *Collaborative learning: Higher education, interdependence, and the authority of knowledge*. Johns Hopkins University Press, 2715 North Charles Street, Baltimore, MD 21218-4363.

Grover, S., Pea, R. & Cooper, S. (2014). Promoting active learning & leveraging dashboards for curriculum assessment in an OpenEdX introductory CS course for middle school. *Proceedings of the first ACM conference on Learning@ scale conference*, 205–206.

Holman, C., Aguilar, S. & Fishman, B. (2013). GradeCraft: what can we learn from a game-inspired learning management system? *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 260–264.

Ososky, S., Brawner, K., Goldberg, B. & Sottilare, R. (2016, September). GIFT Cloud Improving Usability of Adaptive Tutor Authoring Tools within a Web-based Application. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, No. 1, pp. 1389–1393). SAGE Publications.

Sabin, M. (2012). Student-pull instead of instructor-push: in preparation for a student learning dashboard. *Journal of Computing Sciences in Colleges, 27*(6), 70–72.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional science*, *26*(1–2), 113-125.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., . . . Baker, R. (2011). *Open Learning Analytics: an integrated & modularized platform.* Open University Press Doctoral dissertation.

Sinatra, A. M. (2015, August). The Instructor's Guide to GIFT: Recommendations for using GIFT In and Out of the Classroom. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym3)* (p. 149).

Sottilare, R. A., Goldberg, B. S., Brawner, K. W. & Holden, H. K. (2012, December). A modular framework to support the authoring and assessment of adaptive computer-based tutoring systems (CBTS). In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference* (pp. 1–13).

Sottilare, R. A. & Holden, H. K. (2013, July). Motivations for a generalized intelligent framework for tutoring (gift) for authoring, instruction and analysis. In *AIED 2013 Workshops Proceedings* (Vol. 7, p. 1).

Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017. DOI: 10.13140/RG.2.2.12941.54244.

Tucker, B. (2012). The flipped classroom. Education next, 12(1).

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.

CHAPTER 21 – Lessons Learned from Large-Scale E-Assessments: Future Directions for the Generalized Intelligent Framework for Tutoring (GIFT)

Jo-Anne Baird[1], Anne M. Sinatra[2], and Gregory Goodwin[2]
Oxford University Centre for Educational Assessment, UK[1], US Army Research Laboratory[2]

Introduction

Computerized assessment, with all of its promised affordances,[1] has been eagerly anticipated in the field of educational assessment for some time. The technology was embraced in high-stakes testing as soon as it was available, with punch cards being used for statistical processing, optical scanners processing multiple-choice answer sheets, and mainframe computers being brought into service to handle the huge amounts of data processing requirements of running an examination board. Yet, in 2017, we have not seen the revolution in forms of large-scale assessment that we all expected. There are some encouraging, notable developments, which we shall return to, but first, we recount the story so far to understand the issues that may have relevance for the Generalized Intelligent Framework for Tutoring (GIFT) in its venture into the field of computerized educational assessment in conjunction with intelligent tutoring.

Professor Dave Bartrum, Research Director for the SHL Group, former President of the International Test Commission and winner of the British Psychological Society's Award for Distinguished Contributions to Professional Psychology, reviewed the field of automated testing in psychology in 1984 (Bartrum and Bayliss, 1984). He concluded that computerization had helped in generating reports from the data collected through conventional methods. Further, some tests had been transcribed into computerized form without changing the nature of the test content. In such cases, there had been concern about whether this change had brought about a different level of difficulty for test-takers or was simply requiring different skills from a pencil-and-paper testing alternative format. Thus, research on equivalence of these forms began and issues regarding the authenticity of computerized versus paper-and-pencil testing, potential inequalities in opportunities for test-takers relating to computerized testing and the cognitive skills demanded by each form came to the fore.

Additionally, Bartrum and Bayliss (1984) commented that adaptive testing was beginning to be used. In adaptive testing, examinees are presented with an initial question and, depending upon their response, they will be given a question that is (on average) harder or easier. Thus, the test adapts to the apparent level of ability of the examinee. Each test-taker might have a different, personalized test and the duration of the test is shorter because examinees do not have to answer questions that are likely to be far too easy or too difficult for them. As such, adaptive testing was seen to be a very promising development indeed.

Ten years later, the field looked very similar, with generation of reports, equivalence of forms and adaptive testing being the prevalent ongoing developments (Bartrum, 1994). Bartrum made the prediction that within the next ten years we would see the widespread use, understanding, and acceptance of computerized assessment. Few at the time, or at any point in time since, would disagree with this prediction, so the fact that this is not the state of the field requires significant explanation. Surely, the claims that have been made for the potential benefits of technology in this field cannot (all) just be hype, though there is no doubt that there is a lot of over-claiming in this area.

[1] The qualities or properties of an object that define its possible uses.

When it comes to large-scale, high-stakes examinations, the experience for students has been very similar for hundreds (Redecker and Johanssen, 2013), if not thousands, of years. Archives show that in the second century BC, civil service examinations were set in China in much the same way as we experience them today (Lehmann, 2000, 44). So far, as discussed previously, we have seen technology used in assessment to substitute the paper-and-pencil version, with little gain in affordance and there has been some augmentation through the use of adaptive testing (Figure 1). Indeed, technology has been used far more to improve logistics in large-scale assessments than in the user-facing aspects. More adventurous applications, which would significantly redesign the assessment process and even redefine and transform it, are still anticipated (Figure 1). We are on the cusp of moving from reinvention of existing assessments in computer-based assessment to embedded assessments that will transform not only assessment, but learning, according to some (Redecker and Johanssen, 2013). Any time now, we will see the expanded use of automated feedback, behavioral tracking, learning analytics, intelligent tutors, serious games, online collaboration, simulations, and virtual laboratories.

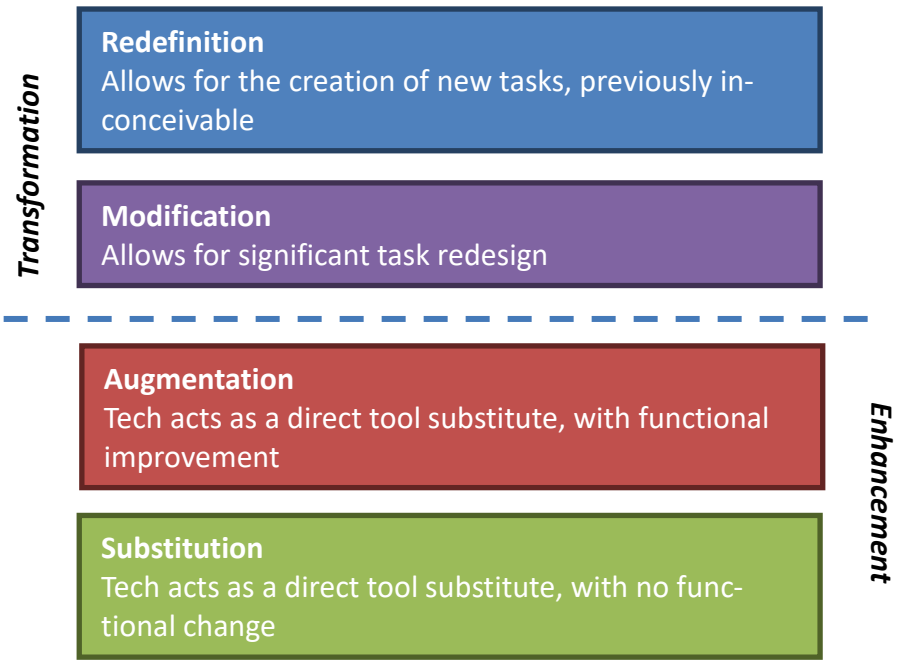


Figure 1. The promise of computerized testing (adapted from Redecker and Johanssen, 2013).

Readers of this chapter are probably coming to this text due to their enthusiasm for these uses of technology. The chapter authors are also enthusiastic, but cognizant of history in this field. GIFT is not in the high-stakes, large-scale testing arena, but perhaps there are some lessons that could be borne in mind before embarking upon assessment design.

Costs of computerized testing have been surprisingly high for many organizations. Software development has not reduced the cost of test production. In the next section, we look at issues related to transparency of assessment and its effects upon the learning process. A transparent, or even a memorable, test is costly because it is unlikely that the same questions can be reused without compromising the validity of the test. Military applications for computerized testing have long spearheaded the field (e.g., the Computerized Adaptive Test version of the Armed Services Vocational Aptitude Battery [CAT-ASVAB]; Segall and Moreno, 1999), so there is good understanding within the community about the likely costs of such an enterprise. Some of the benefits of computerized assessment relate to its capacity to produce relatively authentic contexts for assessment, which should help us to generalize legitimately from test-takers' scores

to their performances in real-life tasks. We discuss the issues related to designing assessments with properties of good generalization. Following this, we look at the distinction between live and simulation training events and the role of the tutor in intelligent assessment. Finally, we conclude with some remarks about how these issues might have applications for the future of GIFT.

Before we turn to those issues, it is useful to first consider to what purpose GIFT assessments might be put. A standard categorization for the possible purposes of educational assessments has been the following:

1. Formative – to aid learning of the student,

2. Diagnostic – to give the tutor information about the student's strengths and weaknesses, and

3. Summative – for decisions regarding review, transfer, or certification of the test-taker.

More recently, the field of educational assessment has come to accept that there are a much wider variety of purposes to which the results of our assessments are put (Newton, 2007). Our point here is not to overly complicate matters, but to flag that assessment design should reflect at least the primary purpose to which the test results are to be put. A test that will license a drone operator will not necessarily have the same design principles as one that is for formative feedback purposes. Many of these issues are subtle and do not always register with subject-matter experts (SMEs), thus assessment design expertise in conjunction with SMEs is an indispensable combination of skills for a team designing GIFT intelligent tutor assessments.

Lessons Learned

Transparency of the Assessment

Large-scale assessments have a history of secrecy, in which the content of the examinations and how the scores or grades were assigned were kept out of the public domain (see Lehmann, 2000). Such secrecy suited commercial interests because the large examining boards could maintain the security of the test items, meaning that they could be reused without fear that people had been able to prepare specifically for those items in advance of the test. Thus, the costs of test development could be kept lower. Further, test producers were authoritative in society and questioning of this status was unwelcome.

If questions can be kept secure, they can be reused without compromising the validity of the assessment. However, if the examinees can remember the questions and there is a high likelihood of them being repeated (high item exposure), even if the questions are kept secure, they could become widely known among the test-taking population. This situation occurred in practice in 1993 when employees of Stanley H. Kaplan Educational Center memorized at least part of 200 GRE questions. This led to the costly, if temporary, suspension of the computer adaptive test. In the legal wrangles that followed, it became the responsibility of test producers to make public their questions and correct answers and examinees have to be able to review the questions they have seen and answers they gave, so that they are not disadvantaged by computer-adaptive methods.

Nowadays there is far more transparency of the test content, scoring, and standard-setting mechanisms even in high-stakes, large-scale tests. Public trust in authority figures has declined in all walks of life and questioning of experts, even doctors, is normal practice (Simpson and Baird, 2013). Tensions are caused by transparency of assessments, as it enables teachers to teach to the test to a larger extent, but even when there was a high level of secrecy, there was always a privileged group who were in the know. Keeping the test content and processes of producing the outcomes secret is not likely to be acceptable to society and makes

it difficult for people to know on what they are being assessed. Secret tests are difficult to prepare for in the right way.

In a computer-based adaptive training system such as GIFT, it is possible to envisage the assessment being conducted by algorithms that operate behind the tutoring system. These algorithms would essentially codify and score learners' behavior as they operate in the system. At one level, this is no different from other assessment systems and the same arguments as discussed above apply. As the algorithms are applied electronically, their operation suffers immediately from a lack of transparency. They are not like examination questions and scoring rubrics, where one can see their operation. Thus, extra caution is required to ensure that they are assessing the right things, rather than spurious behaviors, but ones that can easily be captured electronically. In other words, validity of the assessment depends upon the electronic assessments scoring people's behavior in the right ways. Further, if people are to learn what was intended, it is highly desirable for them to know what will be given credit in the system, so that they can set themselves the right learning goals and work toward them. At worst, electronic assessment systems could become systems that nobody fully understands and that assess invalid aspects of people's behavior. To avoid this situation, it is very important that the construct that is being assessed is spelled out in advance and that there is validation of how the system is measuring that construct (see Chapter 18 entitled, *Validity Issues and Concerns for Technology-Based Performance Assessments*, by Katz et al., *within this volume*). Assessment constructs are usually developed using a mix of four methods:

1) Theory – usually a review is conducted of what is known about the property from empirical research and theory, which shows both its features and what counts as higher and lower values of those properties.

2) Empirical findings on previous tests – often constructs are written with previous test results in mind, as they show what was feasible for learners previously. Additionally, new constructs are often revised in the light of test results.

3) Expert panels – typically, experts in the field are involved in setting out their views of what the construct looks like on the basis of their experience of theory, previous empirical findings and practice.

4) Politics – at some level, whether it be politics with a small or a large p, there are usually aspirations regarding what students must learn that influence the content of the assessments.

Deciding upon the construct that should be assessed is done in conjunction with decisions about what we want to know more broadly about the test-takers; that is, what the scores are likely to tell you more generally about the person's likely performances on other tasks. It is to this issue that we turn next.

Generalization

Interactive tutorial systems like GIFT present the opportunity to collect a wealth of information about individuals' interactions with the system. The output log files from GIFT interactions could be used to generate assessment data. Equally, specific assessments could be set up to collect particular data on individuals' performances. Exactly what data are collected is just as important for validity as the way in which the data are scored. Not only do those data need to tell us about the construct of interest, we want that construct to be some aspect of individuals' performances that is relevant in the real-life task. In other words, we want performances on the GIFT assessment scores to be generalizable to real-world performances. The questions to be addressed when designing the assessment should be about what the designer needs to know about people on the real-world task rather than what data can easily be provided from GIFT. Otherwise, there is

a danger that the technology dominates the design of the assessment and the scores are less informative than might otherwise have been the case.

In designing the assessment, then, consideration needs to be given to the limitations of the simulation. To what extent can GIFT provide data that mimic the real-world task. These same limitations need to be considered when the user of the scores interprets their meaning. Additional assessments might be required that add to the information provided by GIFT to ensure that individuals are properly trained for the working environment. Ideally, a set of valid scores for real-world performances would be available so that they could be used as a criterion with which to validate the GIFT assessment scores. A high correlation between the two sets of scores would provide a justification for generalizing scores from the GIFT assessment to the real-world performances in future.

The foregoing discussed the generalizability of GIFT scores. This relies upon the task performances being generalized to the real-world environment in some way; this is the underlying mechanism. Therefore, design of the tasks themselves need to be carefully thought through in those terms. Some aspects of the task fidelity might not be important for generalization of performances learned in GIFT simulation, but others might be crucial. This is of crucial importance because we know from a wealth of cognitive psychology experiments that people find it very difficult to transfer their training from one circumstance to another (e.g., Evans, 1989). These studies beg the question of how people learn at all if they cannot naturally transfer training from one situation to another. The issue is, of course, that we have to recognize the structure of the problem as being similar and able to apply our learning to that new situation. Scaffolding learning for people, through debriefing for example, is therefore also necessary (Csikszentmihalyi, 1990). Also, the more authentic the training environment, the more likely transfer of training will be. As such, the use of technology to construct more authentic assessment experiences is a very promising feature of GIFT.

Assessments in Live and Simulation Training Events

Military operations typically take place in very complex environments requiring individuals and teams to have a wide range of competencies that they can leverage across a wide spectrum of possible operations. Assessing individuals and units for readiness to deploy to such environments is challenging and requires units to execute complex live and/or simulation based events. Though these are often described as training events, their complexity and cost prohibits multiple iterations and therefore they tend to serve primarily as assessments. That is not to say that all live and simulation-based training serves only to assess. There are many uses of these training venues that are used to practice and refine individual and collective skills.

Assessment in complex scenario-based training events, wherever they are executed, has always been difficult. The Army tends to rely on SMEs who observe, control, and provide feedback during these events. Feedback is most typically a binary pass/fail rating (or in Army parlance go/no-go). Another common rating scale is the three-level trained, practice needed, untrained (TPU) rating. While these ratings are used for recordkeeping purposes, another form of feedback is known as the After Action Review (AAR). AARs are performed at the conclusion of the event and are led by the SMEs who observed the training. They employ a Socratic approach, asking the trainees questions about their actions while guiding them to understand their mistakes. This encourages the trainees to reflect on their actions and decisions, their consequences, and how they as trainees could have done better.

As can be seen, there are summative evaluations (go/no-go or TPU) as well as more formative evaluations (AAR) that take place within scenario based training events. The summative evaluations are more typical of the kinds of evaluations in high-stakes testing, which usually seek to provide a single value that serves to rank the individual on some dimension like quantitative or verbal ability. As with high-stakes testing, these evaluations are part of the unit's or individual's permanent record. Formative evaluations that come

out of the AAR are meant to help the unit and/or individual understand why it succeeded or failed. This feedback identifies specific training needs which drives future training events. There is rarely any permanent, official record of these formative assessments.

As with high-stakes testing, units and individuals are under pressure to receive passing scores on these kinds of assessments. Furthermore, unit leaders are generally responsible for scheduling training and evaluating the performance of their respective units. The advantage of the off-record AAR is that individuals can be much more candid about their failings and can provide constructive ways to improve.

A big challenge of having a system like GIFT assuming a larger role in the evaluation of individuals and units is that leaders will lose some control over those evaluations. As we know from high-stakes testing in public school systems, this process can have some deleterious side effects. For example, leaders will look for ways to game the system to ensure that they receive passing scores. They may begin narrowly teaching and training to the assessment. This may result in leaders neglecting training on skills that, though important, are not part of the assessment.

GIFT and systems like it are really designed to provide very granular formative types of assessments, which drive both feedback and adaptation of the training to specific individuals and teams. Because these assessments are automated, there is the potential for all of these measures to be permanently stored somewhere, if not in a personnel record system. In addition to GIFT's assessments, trainee self-assessments might be recorded by GIFT and used by it during an automated AAR. One has to be concerned about the potential impact of this on the willingness of unit members to be completely candid about their failings during an AAR guided by an automated system like GIFT.

Transparency of measures is also critical. Unit leaders are likely to reject any summative assessment from an automated system unless they can see exactly how that assessment was derived. Furthermore, if they disagree with those assessments, they will lose faith in GIFT. Even at the individual trainee level, transparency of measures will be needed to ensure trust in the assessment.

Though the use of GIFT in these high-stakes scenario based testing events has some potential risks, it also has many benefits. For example, a much richer set of assessments than simply pass/fail is possible. These assessments would be useful to multiple communities including leaders, resource managers, training developers, personnel managers, as well as the trainees. To realize these benefits, it will be necessary to consider how these assessments will be used and by whom as well as how such assessments will be stored and finally how to ensure necessary levels of both transparency and anonymity of these assessments so that effective training can take place.

Role of the Tutor in E-Assessment

While assessment is generally an end goal of instruction, by using computer-based learning there are different techniques or methods that can be used to tailor learning and assessments to an individual. Intelligent tutoring systems (ITSs) allow a learner to engage with material, and provide specific feedback to them based on their knowledge, skills, and abilities. The ITS also can assess the learner's current state and make determinations on material that would be relevant to the learner. In an ITS framework such as GIFT, there is the possibility of going a step further than just tutoring, but to use the given tools to construct assessments that are relevant to the specific learner based on their experiences with the system. GIFT could be leveraged for large-scale assessment in many forms, including, but not limited to, standardized testing and military training scenarios. GIFT authors can either create a static multiple-choice test that always provides the same questions or a dynamic test can be generated for the individual. GIFT allows an author to create a course specific questionbank that can be used by the instructor or test designer to enter all questions that they have constructed or believe are relevant to a given topic. The questions can be tagged with metadata by difficulty

level (novice, journeyman, expert) and also for the concept that they are assessing. The concepts are set up prior to instruction by the course author. The course author can create assessments by determining the number of difficulty level questions that they would like to use for each of the concepts. The specified number of questions would then be randomly selected from the authored questionbank. The questions that were given to the students would be accessible through retrieving the GIFT log files after the fact; however, each test would likely be unique to the specific learner.

If remediation is desired, GIFT's Engine for Management of Adaptive Pedagogy (EMAP) could be used (Wang-Costello, Goldberg, Tarr, Cintron & Jiang, 2013). The EMAP is based on component display theory (Merrill, 1983) and uses the questionbank. Students engage with two types of presented material: rules and examples. Students then engage with two types of "assessments": recall and practice. The recall section of the experience draws the desired number of questions per concept from the questionbank. Based on the performance of the individual during the recall and practice phases, they are sent back for remediation on the concepts if it is found to be necessary. While this may not be advantageous to use during traditional large-scale assessment tests, it is an opportunity for tutoring to occur in a similar environment as the test will occur in. Further, it allows for targeted instruction based on the questions that the individual is missing. GIFT closely ties courses to concepts, and concepts to specific questions; therefore, there is tractability in the assessment that is being provided in GIFT. Once user roles are fully implemented in GIFT, using the questionbank technique will be highly relevant to instructors, as they will have access to the question editor, but their students will not (Sinatra, 2015). The ability to use a questionbank and difficulty level also increases the flexibility of testing concepts with varying questions, which ensures that an individual experience occurs for the learner.

Implications for GIFT

In the opening paragraph, we mentioned that there have been some high-profile developments in the field of computerized testing with large-scale examinations. The Programme for International Student Assessment (PISA),[2] operated by the Organisation for Economic Cooperation and Development, has managed to introduce computerized testing despite the fact that it operated in 71 jurisdictions for the 2015 round of tests.[3] Of course, not all countries used the computerized version and not all students took the tests on computer. The cooperative problem-solving tests, using avatars and natural language processing were a huge step forward compared with any other testing conducted on this scale. Advances in technology will help large-scale assessment managers to overcome the significant cost and logistics difficulties in making best use of the affordances of computerized testing.

As such, GIFT has been designed with flexibility in mind. That flexibility allows for authors to use the provided tools and create interactions that meet their goals. In a traditional GIFT interaction, it would be used as a means of providing adaptive training in a given area and remediation. However, the functionality in GIFT can also be leveraged to provide large-scale assessments, by using the tools in slightly different ways. In the current EMAP-based adaptive courseflow, information is provided to the learner, followed by quizzes, which then result in remediation on the specific topics that are being taught. One suggestion for a future assessment based feature in GIFT would be that instead of remediation being provided on the topic that was missed, additional quiz questions are given on the same topic. Perhaps the questions can rephrase or test different aspects of the initial concept that resulted in low performance, but it would help to target in on the knowledge of the individual and any misconceptions that the learner might have.

[2] http://www.oecd.org/pisa/test/.
[3] http://www.oecd.org/pisa/aboutpisa/pisa-2015-participants.htm.

In current form, GIFT can be integrated with external computer-based games and feedback can be provided during the practice phase of the adaptive courseflow. This may be important in assessment in less traditional domains; however, in more traditional domains such as math or physics, a demonstration of performance may be the completion of a complex problem. It is recommended that in the future both the recall phase of the adaptive courseflow and the practice phase allow for multiple choice and short answer assessment questions. This would lead to different types of tags and functionality in the questionbanks such that recall-based questions could be more straightforward fact-based questions, whereas practice ones could be conceptual and require the solving of a problem. This modification would allow for the adaptive courseflow to be leveraged in a way that would assist in large-scale e-assessment.

While a future direction of GIFT is setting up user roles, making distinctions between the student and the teacher will be vital to using it for assessments. Additionally, setting up ways to report both the real-time and after-the-fact logs of performance to those who are evaluating the tests is a challenge that has not yet been resolved and will be of great importance. GIFT has many functions that could be leveraged or added to assist in using it for e-assessments, with some small changes in the way that some of the adaptations operate, and determining ways to report results it could be an even more powerful tool for large-scale computer-based assessment.

References

Bartram, D. (1994) Computer-based assessment. In Cooper, C. and Robertson, I.T. (Eds.) *International Review of Industrial and Organisational Psychology*, 9, 31–71.

Bartram, D. & Bayliss, R. (1984) Automated Testing: Past, Present and Future. *Journal of Occupational Psychology*, 57, 221–237.

Csikszentmihalyi M. (1990). Flow: the psychology of optimal experience. New York : Harper & Row.

Evans, J. (1989) *Bias in Human Reasoning: Causes and Consequences.* Hove: Lawrence Erlbaum Associates.

Lehmann, N. (2000) *The big test. the secret history of the American society.* Farrar, Straus and Giroux, New York.

Merrill, M. D. (1983). Component display theory. *Instructional-design theories and models: An overview of their current status*, *1*, 282–333.

Newton, P.E. (2007). Clarifying the purposes of educational assessment, *Assessment in Education: Principles, Policy & Practice*,14:2, 149–170.

Redecker, C. & Johanssen, Ø. (2013) Changing assessment – towards a new assessment paradigm using ICT. *European Journal of Education*, 48, 1, 79–96.

Segall, D.O. and Moreno, K.E. (1999). Development of the computerized adaptive testing version of the armed services vocational aptitude battery. Chapter 3 in Drasgow, F. and Olson-Buchanan, J.B. (Editors), *Innovations in Computerized Testing*, 35–66.

Simpson, L. and Baird, J. (2013) Perceptions of trust in public examinations. *Oxford Review of Education*, 39, 1, 17–35.

Sinatra, A. M. (2015, August). The instructor's guide to GIFT: recommendations for using GIFT in and out of the classroom. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym3)* (p. 149).

Wang-Costello, J., Goldberg, B., Tarr, R.W., Cintron, L.M. & Jiang, H. (2013). Creating an advanced pedagogical model to improve intelligent tutoring technologies. In The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC).

SECTION IV

ASSESSMENT METHODS FOR PARTICULAR DOMAINS AND PROBLEMS

Dr. Xiangen Hu, Ed.

CHAPTER 22 – Selected Assessment Techniques for the Generalized Intelligent Framework for Tutoring (GIFT)

Xiangen Hu
University of Memphis, Central China Normal University

Introduction

In this section, there are a total of seven chapters focusing on various aspects of assessment techniques of intelligent tutoring systems (ITSs). While these seven chapters do not cover all possible assessment techniques in ITSs, they nevertheless offer a good view on a few very important aspects. To understand the contributions of the authors, I suggest the reader first consider the following: 1) the special role an ITS plays in an advanced learning environment (ALE) where the ITS serves the role of an educator, 2) that in the ALE, if any two components interact in an active fashion, there is a need of assessment, and 3) the structure of the General Intelligent Framework for Tutoring (GIFT), which has been introduced and prototyped by the US Army Research Laboratory.

Adapting from a general definition of educational assessment: In education, the term "assessment" refers to the wide variety of methods or tools that educators use to evaluate, measure, and document the academic readiness, learning progress, skill acquisition, or educational needs of students (http://edglossary.org/). This general definition of assessment is applicable for assessing the overall quality of the ALE, except the *educators* here has a double meaning: learning scientists who design and implement the ALE in education systems, and ITSs that serve as an educator (human teacher) that educates students. From this perspective, the assessment techniques offered by the authors consider both the assessment needs for learning scientists to evaluate education systems and the specific needs of computer tutors (ITSs) for assessing learner's knowledge, skills, and attitude.

In these seven chapters, the authors recommend enhancements/extensions of the GIFT assessment techniques in several dimensions. Chapter 23 recommends incorporating assessing GIFT's overall quality/effectiveness based on technology that evaluates the quality of questions for teachers in the classroom. Chapters 24 and 25 recommend an assessment technique for GIFT in team/collaborative tutoring environments. Other authors recommend assessment techniques based on different types of data: Chapter 26 considers discrete and categorical behavior, Chapters 26 and 27 consider natural language interactions, and Chapter 28 examines time latency and accuracy. In addition to the usual assessment of cognitive skills that most of the assessment models/techniques address, Chapter 29 recommends an assessment technique for GIFT to assess the learner's motivation.

Outline by Chapters

In Chapter 23, *Assessing Teacher Questions in Classrooms*, Olney, Kelly, Same, Donnelly, and D'Mello offer their recommendation based their research and development of a system called CLASS 5.0, which automates the process of classroom observation through speech recognition and machine learning. They specifically recommend a technique to assess ITS question-asking behavior. Given the ITS plays the role of a human teacher in an ALE and the evidence that if human teachers ask better questions, their students learn better, they assume that if there are mechanisms for assessing the question-asking behavior of natural language-based ITSs, then the overall quality of an ITS can be evaluated. Notice the goal of implementing this assessment is not assisting the ITS to assess the learner, instead this recommendation suggests an assessment technique for the overall quality control of an implemented GIFT application (of a given domain).

In Chapter 24, *Assessment of Collaborative Problem Solving*, Graesser, Cai, Hu, Foltz, Greiff, Kuo, Liao, and Shaffer introduce a Collaborative Problem-Solving Framework (CPSF) from PISA 2015 and propose a possible enhancement to the current GIFT in the following steps: 1) extending GIFT to handle team ITSs and 2) consider CPSF as an assessment framework evaluating students' collaborative ability for the extended GIFT. Successful implementation of the two steps for an extended GIFT will make it possible to 1) analyze of interactive log data to assess the learner as a single team member or assess the performance of a mixed team of learners and GIFT, and 2) use the CPSF as a recommending system (such as production rules) to guide the tutoring behavior of the extended GIFT.

In Chapter 25, *Challenges for Assessing and Tutoring Collective Skills*, Ayers, Bink, and Diedrich provide three recommendations based on their research and development (R&D) in systems that involve coordination and communication between human and/or synthetic actors. It is understandable that such an ALE would be more complicated than the existing ITSs that have been implemented where the ITS spends most of the time interacting with single learners.

In Chapter 26, *Cognitive Assessment as Service in the General Intelligent Framework for Tutoring (GIFT)*, Hu, Xu, Sottilare, and Albert recommend a possible assessment service component for GIFT. They demonstrate the feasibility for such a complement by providing two sample assessment web services. One is the Multinomial Processing Tree (MPT) model, which is in the field of cognitive psychometrics, and the other is Semantic Representation & Analysis (SRA). The MPT service can be used to assess a student's knowledge by analyzing categorical behavior and the SRA service to assess a student's knowledge by analyzing natural language interactions with the ITS.

In Chapter 27, *Assessment in AutoTutor,* Cai, Graesser, Hu, and Kuo provide an assessment technique based on the current implementation of the AutoTutor framework. Their recommended assessment techniques includes the assessment of a student's knowledge in an inner loop (within AutoTutor) and an outer loop (outside of AutoTutor). The recommended technique in the inner loop (of AutoTutor) uses semantic analysis technique to measure the similarity between students' verbal contributions and pre-stored semantic answers. The semantic analytic methods have proven efficient in assessing students' knowledge and can be used to create a learner model. The recommended technique in the outer loop uses another research-based knowledge organization/domain modeling technique, Learning Space Theory.

In Chapter 28, *Assessment of Individual Learner Performance in Psychomotor Domains*, Kim, Sottilare, and Goodwin report their previous research and implementation of a three-stage model of a *theory of skill learning and retention.* Their R&D indicates that it is possible to model psychomotor domain with this theory. They recommend a possible extension of GIFT to include sense/location aware devices such as smartphones to assess trainees' knowledge from the speed and accuracy information on psychomotor tasks.

In Chapter 29, *Motivating Individual Difference in an Intelligent Tutoring System*, Reinerman-Jones, Lameier, Biddle, and Boyce provide a general framework for assessing learners' motivational differences. Most importantly, they introduced a framework that includes six factors that influence learner motivation. They go further by recommending methods for tool development and strategies for implementation.

CHAPTER 23 – Assessing Teacher Questions in Classrooms

Andrew M. Olney[1], Sean Kelly[2], Borhan Samei[1], Patrick Donnelly[3], and Sidney K. D'Mello[3]
University of Memphis[1], University of Pittsburgh[2], University of Notre Dame[3]

Introduction

Prompted first by the federal No Child Left Behind Act and subsequently the Race to the Top grant program, states have moved to adopt accountability systems that not only hold schools accountable for producing learning, but also individual teachers (Gamoran, 2013; Kelly, 2012). These efforts are consistent with research demonstrating the variability in teachers' capacity to improve student achievement growth (Hanushek & Rivkin, 2006, 2010; Kane et al., 2013; Nye, Konstantopoulos & Hedges 2004). In most cases, educator evaluations are based in part on student test scores, but also incorporate observational measures of instruction. Observations of classroom practice are valuable because they capture dimensions of schooling not captured by test scores, such as socialization outcomes in elementary school (Jennings & Corcoran, 2012). Classroom observations also enhance school principals' role in managing teachers' work (Harris, Ingle & Rutledge, 2014). Moreover, the presence of observational measures places an emphasis on the process of instruction itself and can be used to facilitate professional development quite apart from teacher evaluation (Goe, Biggers, Croft, 2012). Thus, many experts advocate balanced systems of accountability that include observational measures of instruction (Gates Foundation, 2013; Hamilton, 2012; Stein & Matsumura, 2009).

To date, several observational protocols have been developed, some for use across multiple classroom contexts (e.g., the Danielson Framework for Teaching [FFT], see Sartain, Stoelinga & Brown, 2011), and some targeted to instruction in specific subjects (e.g., the Protocol for Language Arts Teaching Observation [PLATO], see Grossman et al., 2013; and the Mathematical Quality of Instruction Instrument [MQI], see Hill et al., 2008). Systems of observational evaluation are currently in use in 47 states in the United States (American Institutes for Research, 2016). Yet, current methods are logistically complex, requiring observer training and are also an expensive allocation of administrator's time (Archer et al., 2016). For example, for use in evaluation, studies show that, typically, four class observations of each teacher are needed to provide a reliable sampling of teachers' instruction and afford an adequate opportunity to demonstrate excellence in multiple instructional domains (Kane & Staiger, 2012). Without a carefully managed classroom observation process, the observation results are open to criticisms of bias or arbitrariness.

To address the problems of cost, reliability, and bias inherent in traditional observational protocols, we have undertaken the development of a system called Classroom Language Assessment System (CLASS) 5.0 that automates the process of classroom observation through speech recognition and machine learning. The primary focus of our work is on teacher question-asking behavior, which is a common component across various well-known observation protocols. We first review some recent results in the classroom observation literature before describing our own work and making recommendations for future research.

Observing Effective Teaching

Efforts to evaluate teachers' performance are based on the logical assumption that the individual teacher's classroom is the most important site of student learning, and thus, closer evaluation and support for teachers' work constitutes a powerful lever of educational reform. Basic research on teacher effectiveness supports this perspective (Hanushek & Rivkin, 2010; Kane et al., 2013; Konstantopolous, 2014). For example, Nye et al. (2004) estimate that a 1 standard deviation increase in teacher effectiveness would increase student achievement by about 1/3 of a standard deviation. Other research finds somewhat smaller, but still

important achievement gains attributable to teacher-to-teacher variability (Cantrell & Kain, 2013; see Hill et al., 2008 for a discussion of interpreting effect sizes in education research). In contrast, in much prior research, readily available indicators of teacher quality, such as years of experience, educational attainment, or certification status have generally explained a frustratingly small proportion of the variance in teacher effectiveness (Clotfelter, Ladd & Vigdor, 2006; Hanushek, 1986). Given these two sets of findings, directly assessing individual teachers' performance via test scores and/or teacher observations may offer the best insight into teaching quality. What does existing research demonstrate about the role of observation in assessing effective teaching?

A recent, multi-year study called Measures of Effective Teaching (MET) examined several different classroom observation measures, student perception surveys, and student achievement gains across approximately 3,000 teachers in seven states (Cantrell & Kane, 2013). Although previous research has shown a connection between various classroom observation measures and student achievement, the MET study is unusual in two respects. First, it used a randomized controlled trial design that, in year 1, collected teaching effectiveness data and built predictive models of teaching effectiveness, and in year 2, randomly assigned teachers to new classrooms to see if the predictive models from year 1 could account for changes in student outcomes in the randomly assigned classrooms. The purpose of the randomized controlled trial was to establish a causal, rather than correlational, correspondence between teaching quality and student outcomes, making MET the largest study of its kind to do so. Second, classroom observations were conducted using recorded video, allowing multiple observers and diverse observation measures for each video. This approach allowed for an in-depth examination of the reliability of the various classroom observation protocols across different types of observers.

MET used both general and subject specific classroom observation protocols, such that various observational measures of effectiveness could be compared to each other and value-added estimates of student achievement growth (Mihaly et al., 2013). One of the major MET findings was that these classroom observation protocols were all positively correlated with student achievement gains and were highly correlated with each other at the summary score level when using dis-attenuated correlations to account for measurement error (Kane & Staiger, 2012). Considering the correspondence in teacher ratings across different observational protocols, FFT, Classroom Assessment Scoring System (CLASS), and PLATO had pairwise correlations above 0.86, and FFT, CLASS, and MQI had pairwise correlations above 0.67, ostensibly lower due to the specific mathematics focus of MQI. A principal component analysis on each protocol yielded three major components that accounted for approximately 90% of the variance in scores across teachers. The first two components were the same across protocols: overall quality and classroom management (Kane & Staiger, 2012). The third factor varied across protocols, but most often involved question asking behavior, the focus of this chapter. Considering the relationship between observational scores and achievement gains, the teachers rated most effective on observations were also effective in raising test scores. For example, correlations between FFT (Danielson, 2011) scores and the value-added achievement measures (state tests) ranged from 0.17 to 0.41 (Mihaly et al., 2013, Table 3).

Measurement of Dialogic Instruction

One limitation of the protocols used in the MET study is the relatively coarse-grained nature of the coding as CLASS, FFT, and PLATO were all coded on 15-minute intervals, i.e., every 15 minutes, while MQI was coded on 7.5-minute intervals (Kane, Kerr & Pianta, 2014). The time delay between a classroom event and the coding of it on these interval boundaries may have facilitated observer's use of holistic judgments of instructional quality, which would explain the lack of differentiation among the various dimensions of these protocols with respect to the first principal component (e.g., the FFT has 22 rating components and 76 smaller elements within them).

In contrast, the present study is based on Nystrand's CLASS, a real-time, or "live" coding system first developed by Martin Nystrand and colleagues in the mid-1980s (Nystrand, 1988). Nystrand's CLASS focuses on individual questions and their properties, in addition to the basic allocation of classroom time to various instructional activities. In this study, we use data from updated versions of the original CLASS program, which was used in coding both archival data from the Partnership for Literacy Study (see Kelly, 2008) and in newly collected data. We pair these human coded measures from the CLASS program with new automated codes, referring to the automated version of the system as CLASS 5.0. Note that the MET study used a separate system, also called CLASS, developed by Robert Pianta and colleagues (see Allen et al., 2013).

The micro (i.e., individual question events) rather than macro orientation of CLASS 5.0 is highly salient to adopting a machine learning approach to classroom observation, because it provides labeled data conducive to training classification models. Application of machine learning to the data coded in the MET study seems much less promising, because the long durations between class events and actual coding creates a credit assignment problem (Minsky, 1961) in which it is unclear what action or event led to a given code. In CLASS 5.0, the relationship is much more direct, though not perfectly so. Instructional segments (e.g., discussion, lecture) have clear timestamped boundaries and categories. Within these timestamps, classroom activities and language are strong markers of the segment category. Questions likewise have clear timestamps, as do some clear properties like speaker identity. However, some associated properties extend beyond the question per se and instead are more properly considered part of a question event. These question properties include whether there was a response, cognitive level, authenticity, and uptake. These last two properties are the hallmarks of dialogic instruction, in which questions do not have predefined responses (authenticity) and are part of an evolving discussion that incorporate ideas from the respondent (uptake). Compared to common initiation-response-evaluation (IRE) format of classroom instruction in which the teacher quizzes students by asking them "test" questions, dialogic instruction focuses on the open-ended discussion and the exchange of ideas (cf. Bakhtin, 1981). For example, "What was your reaction to the end of the story?" is an authentic question because there is no pre-scripted response, and a follow-up question "Why do you think that?" has uptake because "that" refers to the student's previous reply. As is clear in these examples, dialogic properties are contextualized by the discourse and not purely determined by the question alone. Thus, the question event is characterized by the antecedents and consequents of the question in addition to the question itself.

Currently, it is not clear from the MET results whether instructional processes surrounding question-asking behaviors (one of the principal components of effective instruction in the MET study) have a significant effect on student achievement. On the other hand, previous research on the CLASS 5.0 system has shown that authenticity and uptake are significant predictors of student achievement (Gamoran & Nystrand, 1991; Gamoran & Kelly, 2003; Nystrand & Gamoran, 1997). Moreover, teacher training can increase the prevalence of dialogic instruction (Caughlan, Juzwik, Borsheim-Black, Kelly & Fine, 2013). Thus, the rationale of our work is that dialogic instruction, by promoting student achievement and being responsive to professional development, might be a crucial factor to assess when it comes to using classroom observation for measuring teaching effectiveness. In addition, unlike principal components of overall quality derived from statistical analyses of covariance, it can be precisely defined. However, an obvious limitation is that question-asking behaviors constitute a narrower domain of classroom instruction than assessed in global ratings. For example, in English and language arts, many important instructional dimensions (e.g., goal clarity, challenge) pertain to writing activities rather than discourse. Overall though, we view the measurement of dialogic instruction as an appropriate target of classroom observation to improve teaching effectiveness via feedback and professional development.

Class 5.0

In our work, we have pursued the goal of measuring dialogic questions in classrooms from two perspectives, text based and speech based. Text-based work makes use of both human transcripts of questions, which are available as archival data from previous CLASS 5.0 research, and automatic speech recognition (ASR) transcripts derived from current data collection. Speech recognition in classrooms is quite challenging given the ambient noise and the impracticality of individual microphones for each student. As a result, we have primarily focused on high-quality teacher audio collection and lower-quality classroom audio collection, from which we can determine student speech activity (D'Mello et al., 2015). Although speech features like prosodic, spectral, and voice quality features do contribute to the accuracy of question detection and instructional segment classification, text-based features alone are very effective (Blanchard, D'Mello, Olney & Nystrand, 2015; Blanchard et al., 2016b, 2016a; Donnelly et al., 2016a, 2016b, in press). Thus, text-based features are central to assessing classroom discourse, so there is considerable need for transcripts from which to extract these features. Perhaps surprisingly given the recent advances in ASR and industry claims of word error rates at 10% or less (Ryan, 2016), word error rates in our classrooms are closer to 50%, even when using a high-quality microphone and available state-of-the-art deep neural network-based ASR systems (Donnelly et al., in press).

We are currently still exploring how speech recognition errors differentially affect our models at the feature level. In our previous work, based on archived human transcripts from 418 classes, we trained J48 decision trees (Quinlan, 1993) that were able to automatically detect dialogic question properties at approximately 64% accuracy, which rivals humans coding questions out of context (Samei et al., 2014). In contrast, using new data collection ASR transcripts from 77 classes, our models are only 54% accurate. There are two likely causes for this difference in performance between models trained on human transcripts and ASR transcripts. First, ASR errors could be corrupting key features needed to build successful models. Second, the new data collection ASR transcripts have only a fraction of the original amount of human transcript training data, so the difference could be that there is not enough data to build a successful model. Both likely causes for performance differences may be examined by subsampling the human transcripts, i.e., creating new data sets for training by sampling without replacement various percentages of the total data set. Figure 1 shows model performance on human transcripts at 1% to 100% increments of the total data set. As the amount of data increases, model performance (defined as percent correct) improves, but the growth of improvement slows as more data are added.

In terms data size, the amount that is available from new data collection with ASR transcripts is approximately 6% of the human transcript data set size, for which Figure 1 shows trained models are 57.6% percent correct. Based on this analysis, it seems likely that a significant part of the difference between human transcript and ASR transcript model performance is due to lack of data, because when there is a comparable lack of data in the human transcript data set, the results are only 3.6% better rather than 10% better as they initially appeared to be. Overall, this implies that ASR errors are only negatively impacting model accuracy by 3.6%, and that this deficit could be narrowed simply by collecting more ASR transcript data for training. Using the equation of the line of best fit in Figure 1, the amount of data required would be approximately 2.5 times the number of human transcripts currently available. However, this estimate should be treated with caution since there is no guarantee that ASR based models will exactly follow the curve for models trained on human transcripts.

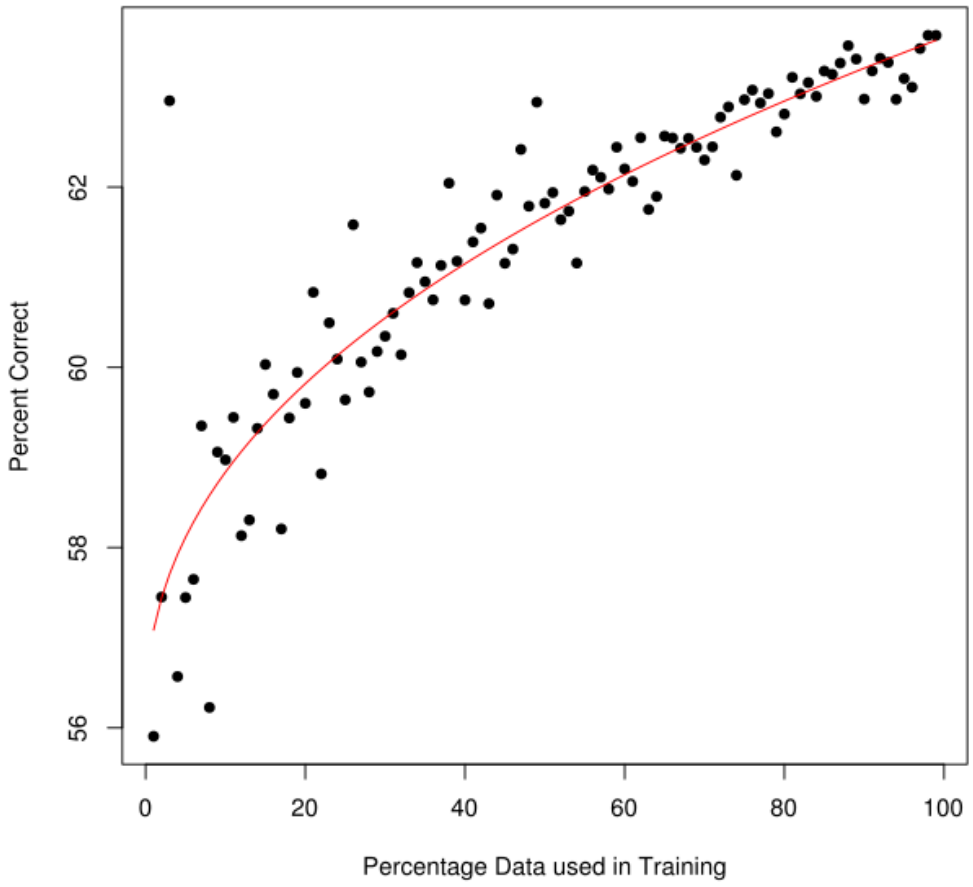


Figure 1. Model accuracy as a function of training data size for models trained on human transcripts.

In this chapter, we focus on our recent work training models for authenticity and uptake using human transcripts of questions collected in previous projects by Nystrand and colleagues.

Assessing Questions in Isolation

Our initial work on the assessment of dialogic questions focused on questions in isolation, meaning questions removed from the discourse context in which they occurred. The rationale for this line of inquiry was pragmatic, in terms of both the data available and building better machine learning models. The data available at the beginning of our project consisted entirely of archival data produced by previous versions of CLASS 5.0. These data contained human transcriptions only of coded questions and not of the surrounding speech. Non-instructional questions, such as procedural questions, were excluded from the coding scheme (Nystrand, 1988).

From the perspective of creating machine learning models, investigating isolated questions is also pragmatic because it raises the question of just how much information is needed to measure dialogic questions effectively. The human observers, situated in the classroom during live coding, have access to a considerable amount of contextual information, including spoken language, non-verbal communication, whether

the students are paying attention, etc. Although we assume that the bulk of the coding decisions are based on spoken language, it may be the case that these other sources of information have some role to play.

Therefore, one of our first questions was whether a new set of trained human coders would code isolated questions as accurately as the original live coders. We randomly sampled 200 questions for authenticity and uptake (400 in all) from the approximately 25,000 questions in our archival data. The questions were evenly balanced such that half had the property in question, e.g., uptake, and the other half did not. Four trained human judges recoded these questions independently, with questions being presented in a random order. For comparison, though studies from CLASS 5.0 prehistory did not use chance-corrected agreement statistics like Cohen's kappa, inter-rater agreement defined as percent agreement has been reported as 81.7% for uptake and 78% for authenticity (See Nystrand & Gamoran, 1997, Chapter 2, Footnote 3).

Inter-rater agreement for the authenticity and uptake samples is shown in Tables 1 and 2, respectively, which are elaborated versions previously published in Samei et al. (2014). Three patterns are apparent in these results. First, kappa between new raters (R1-R4) on the isolated questions ranges between 0.18 and 0.46 for authenticity and between 0.31 and 0.51 for uptake. Although interrater reliability appears to be low, it corresponds to historical percent agreement on this task, given that 80% agreement on two evenly balanced classes would yield a kappa of about 0.35 (Bakeman, McArthur, Quera & Robinson, 1997). Thus, the agreement between new raters is reasonable except for the low 0.18 kappa between R3 and R4 for authenticity. Second, the kappa between the new coders and the original (O) live coders drops substantially and is equivalent to a 20–30% drop in percent agreement. It is noteworthy that while the original live coders presumably agreed with each other at the same level as the new coders, agreement between original and new coders is low. This indicates that different criteria are being used by original and new coders as the basis for their coding decisions. Third, the kappa between the J48 decision tree model (M) and the original live coders is substantially higher than the kappa between the original live and new coders. Indeed, the kappas between the model and the original live coders approaches what we would have expected to see between the original live coders themselves, if that data were available.

Table 1. Kappa for authenticity on isolated questions.

Rater	R1	R2	R3	R4	M
R2	0.44	-	-	-	-
R3	0.41	0.36	-	-	-
R4	0.46	0.55	0.18	-	-
O	0.13	0.17	0.25	0.10	0.34

Table 2. Kappa for uptake on isolated questions.

Rater	R1	R2	R3	R4	M
R2	0.45	-	-	-	-
R3	0.31	0.46	-	-	-
R4	0.51	0.47	0.36	-	-
O	0.22	0.25	0.30	0.23	0.46

As further discussed in Samei et al. (2014), these results are somewhat tempered by the finding that when considering the entire data set (i.e., approximately 25,000 questions), the percent agreement between the model and the original coders is 64% for authenticity and 62% for uptake. Thus, it appears that our models

still need to account for 15–20% agreement to be on par with live coders, at least on this task where human transcripts of questions are classified without any context.

Assessing Questions in Sequence

We repeated the recoding task in various forms, incrementally including information such as who was speaking (useful since student questions are more likely to be authentic than teacher questions). In our latest evaluation, we think we uncovered the simplest combination of factors that can be used in a machine learning model. These include speaker identity and question transcript for all questions in each question/answer segment (assuming instructional segment boundary detection and classification). Two raters each with over 10 years of experience in coding dialogic questions independently rated a sample of 102 questions. The questions were sampled at the segment level, meaning question/answer segments were first randomly sampled, and then all questions from these segments were extracted in temporal order. The raters were presented with lists of questions with corresponding speaker identity (either teacher or student), question transcript, and segment boundaries. There were 14 segments in all, ranging from 1 to 24 questions in length. The prevalence of authenticity and uptake were representative of the entire dataset, with 48% of questions being authentic and 30% of questions having uptake. Unlike previous work, these properties were coded simultaneously for each question rather than having separate question sets for each.

The inter-rater agreements in kappa are shown in Tables 3 and 4 for authenticity and uptake, respectively. It should be noted that one coder (R1) failed to code six questions coded by both R2 and the original live coder (O). Therefore, kappas involving R1 are only based on the 96 questions that were rated, but all other kappas are based on the full set of 102 questions. Given that the kappa between R1 and O, using only 96 questions, is only 0.03 higher than the kappa between R2 and O, using all 102 questions, the exclusion of six questions appears to be contributing a very small bias in agreement, if any. Unlike the results for isolated questions in Tables 1 and 2, agreement between the new coders and the original live coder was quite high; in one case (R1 to O) the kappa of 0.61 is quite high for authenticity. In terms of percent agreement, these results range from 72% to 81% for authenticity and 71% to 77% for uptake. Although different coders achieved the highest agreements for authenticity (R2) and uptake (R1), the fact that they were able to do so with this restricted set of information suggests that it is possible for a machine learning approach to do equally well given the same information. This is remarkable considering that dialogic discourse is defined by the antecedents and consequents of questions, as only this context reveals the function, or effect, of a given question on the discourse. However, it appears to be the case that speaker, question transcript, and segment identification convey the same information or sufficiently correlated information to perform the coding task as well as a live coder with full access to the classroom context. Accordingly, building models with only this information is a focus of our current work.

Table 3. Kappa for authenticity with identity and segment information.

Rater	R1	R2
R2	0.48	-
O	0.44	0.61

Table 4. Kappa for uptake with identity and segment information.

Rater	R1	R2
R2	0.31	-
O	0.45	0.41

Assessing Questions without Bias

Although accurate measurement of classroom discourse is of considerable importance, of equal importance is unbiased assessment with respect to various socio-economic factors. This is a growing concern in artificial intelligence research (Hardt, Price & Srebro, 2016) because data-driven models will naturally reflect the biases in that data. When the predictions of the model are heavily weighted in high-stakes decisions, such as teacher assessment for promotion or tenure, it is critically important to ensure that all teachers are treated fairly. To better understand how our models were affected by bias, or equivalently to demonstrate that they work equally well for various socio-economic groups, we undertook an analysis of our original work that measured the dialogic properties of question in isolation. Specifically, we subdivided the data for various groups, built models with those subsets, and tested those models against different subsets as well as the full data set (Samei et al., 2015).

The distribution of schools in different geographic regions is shown in Table 5. Because the amount of data in some of these schools was rather small, we grouped them into Urban (Mid-size and Large Central City) and Non-urban groups (everything else). Furthermore, we were able to divide the data into groups who had received professional development training on dialogic instruction (Post-training) and those who had not but would later (Pre-training).

Table 5. Distribution of schools by geographic area.

School Category	Schools	Schools (%)
Large central city	4	19
Mid-size central city	7	33
Urban fringe of mid-size city	7	33
Small town	1	5
Rural inside MSA*	1	5
Rural outside MSA*	1	5

* Metropolitan Statistical Area.

The new groups we defined had different levels of uptake and authenticity, as shown in Table 6. Non-urban and Post-training groups had higher levels of authenticity and uptake than Urban and Pre-training groups. Also, the difference between Pre- and Post-training levels of authenticity and uptake was relatively large compared to the difference between Urban and Non-Urban levels. These different levels of authenticity and uptake across groups suggest the potential for bias if data from one group were used to build a model for another group.

Table 6. Percentage of authenticity and uptake across groups.

Group	Authenticity (%)	Uptake (%)
Non-urban	54	23
Urban	47	20
Post-training	52	24
Pre-training	39	15
Full	50	21

Note. Adapted from Samei et al. (2015).

To investigate the possibility of bias, we built two kinds of models for each group. In tenfold cross validation models, we used each group for both training and testing data. The second set of models trained on a group and tested on its dual, i.e., Pre-training vs. Post-training and Urban vs. Non-urban. The accuracies of these fitted models are shown in Table 7. For comparison, Table 7 also shows tenfold cross validation on the full model.

Table 7. Accuracy of models for authenticity and uptake when trained and tested on different groups.

Training Data	Test Data	Authenticity Accuracy (%)	Uptake Accuracy (%)
Non-urban	Non-urban	61	59
Urban	Non-urban	62	62
Non-urban	Urban	60	63
Urban	Urban	62	60
Post-training	Post-training	63	61
Pre-training	Post-training	59	62
Post-training	Pre-training	60	64
Pre-training	Pre-training	64	61
Full	Full	64	62

Note. Adapted from Samei et al. (2015).

As shown in Table 7, training with Urban gives better results for authenticity than does training with Non-urban regardless of which group is used as test data, while training with Non-urban gives better results for uptake in the case of testing with Urban only. Pre- and Post-training give best results for authenticity when trained and tested against themselves (using tenfold cross validation), but give best results for uptake when tested against the other. These inconsistent results suggest that using any one group to train will create bias of some kind when testing using another group.

To investigate these inconsistent results, we analyzed the individual features used in each model using the Correlation-Based Feature Subset (CFS) algorithm. We found that different subgroups used different kinds of language to mark authenticity and uptake. For example, Urban groups used first and second person "be" verbs and judgmental words like "think," "find," and "thought" in authentic questions, but Non-urban groups did not. Likewise, Urban groups used second person pronouns and negation in questions with uptake, but Non-urban groups did not. Post-training groups had a greater prevalence of "be" verbs for authenticity, and a greater use of modal verbs like "would," "can," and "could," than did Pre-training groups.

For authenticity, training on the full data set led to the better or equal performance than training on any subset. We speculate this is because all subgroup-specific features were represented in the model and prevented it from being biased toward or against dialogic classifications because of the absence of a diagnostic feature. For uptake, training with the full data set is slightly worse than training with Non-urban or Post-training and testing on their duals. However, when tested on themselves (tenfold cross validation), both Non-urban and Post-training are worse than training and testing on the full data set.

While the best possible scenario would be for there to be no difference between groups, the current result can be considered the second best: there are differences, but they can be modeled without explicitly defining the groups in the model. A worse scenario would be if different groups used language in opposite ways, i.e., a marker for authenticity in Urban subset was a marker for non-authenticity in Non-urban subset. If this were the case, then it would be necessary to infer which group was being measured to "select" the right markers for measurement. Fortunately, it appears that group identity can be ignored at the modeling stage

if the training data are sufficiently diverse to represent all groups. Diversity in the training data is critical for unbiased assessment in our models.

Conclusions and Recommendations for Future Research

We reviewed current work on classroom observation in the study of teacher effectiveness as well as our own work on measuring the dialogic properties of questions in classrooms. Previous research, including the MET project, has shown that though the year-to-year effects of high instructional quality relative to lower quality instruction are sometimes small, the cumulative effects across a student's K–12 career can be considerable. Moreover, classroom observations can be used to reliably identify effective instructional practice, and form the basis of professional development efforts and other approaches to school reform. Our work shows that automation holds much promise in scaling up classroom observations in a reliable and fair way.

However, several questions raised by our work present a challenge to future researchers. Currently, we have compared automated coding to relatively fine-grained, question-level human coding, but much existing educational improvement efforts use even courser-grained, global rubrics. Thus, further research is needed comparing automated portraits of effective instruction to a greater array of human-coded approaches. Might the MET data constitute a promising existing resource for such analyses? One concern is that the quality of the audio in these recordings is challenging, and there is also a practical barrier because the MET videos can only be accessed via a remote interface that includes a video viewer, but nothing else. Nevertheless, the potential is great given the large numbers of videos and overall diversity of the MET sample.

A second question for automation in future research concerns the connection between the dimensions of instruction that can be observed *by discourse alone* and achievement growth. In MET, the principal component analysis did identify a component associated with discourse. Yet, further research is needed, building on the MET design, to understand the robustness and malleability of discourse effects in contrast to more generic domains of practice (e.g., that include writing assignments).

Third, the archival data used in this study date to the early days of NCLB and, and it is possible that the prevalence of dialogic discourse and teacher-to-teacher variability in approaches to discourse have changed. On the one-hand, increased teacher accountability and other standards-based reforms may focus attention on test preparation activities and away from dialogic approaches. On the other hand, effective discourse practices, including dialogic practices, are an explicit component of common observational protocols used to evaluate teachers. Are current trends in education promoting teaching practices that are consistent with what research deems effective? Analyses of the MET and other new data may shed light on these questions.

Another question, closely following our own work, is how to build models based on the finding that speaker identification, question transcript, and segment identification seem to be all that is needed to reach live human-coder levels of agreement for dialogic question properties. A related question pertains to whether model-coded dialogic question properties predict achievement gains as well as human codes. If successful, the CLASS 5.0 system can be used a tool for accurately coding classroom discourse, thereby providing valuable information to researchers, teachers, teacher educators, and professional development personnel.

This work is relevant to GIFT in at least two ways. First, dialogic questions could be incorporated into intelligent tutoring systems (ITSs). The work reviewed in this chapter provides a foundation for the generation of such questions computationally. In some respects, this is trivial: asking the user what they think about a topic without some pre-scripted answer or weaving what the user says back into the conversation are common strategies used by chatbots. However, the simplicity of generating such questions is offset by the complexity in keeping the conversation going once they have been asked – in essence, understanding the student's response well enough to generate new dialogue on the fly. This ability, currently lacking in

chatbots, is considered by some to be the ultimate proof of artificial intelligence, the so-called "Turing Test".

Thus, incorporating dialogic questions into ITSs may be beyond the current state of the art, but the existing work in GIFT supporting user personalization provides a starting point from which to build. For example, Sinatra (2015) proposes using a dialogue template approach where the user's log in name is stored as a variable and then inserted into dialogue templates to create dialogue like, "Welcome to the tutorial, [name]" and also proposes a survey to elicit user interests so that they can likewise populate templates for instructional dialogue. These proposals are similar to what currently is done in chatbots using a variable/template approach, but they differ in terms of how the variables are assigned. In the case of GIFT, Sinatra's proposal is to assign these variables outside of the dialogue. GIFT could benefit from variable assignment both outside and inside the dialogue to support dialogic instruction, as internal variable assignment taking place in dialogue would make it easier for an intelligent tutor to weave the student's responses back into the conversation.

The second way in which the work described in this chapter is relevant to GIFT is for hybrid instruction in which the human instructor and GIFT are members of the same instructional team but with different roles. For example, the human instructor may lead a face to face session with students and then pass the students off to GIFT for self-directed practice. In such a scenario, it is important for GIFT to understand what has taken place during the human-led portion of the instruction. Automated assessment of classroom discourse provides such a model of understanding. By using the techniques described in this chapter as well as the techniques by D'Mello and colleagues on instructional segment classification, GIFT could understand both coarse-grained classroom activities and fine-grained discussion, and use this knowledge to tailor its own instructional activities. For example, if the human instructor spent 20 minutes lecturing on a topic to provide an overview, GIFT could trim or eliminate that portion of its instruction. Similarly, highly dialogic discussion during the human-led portion could be modeled and used by GIFT, e.g., in the user personalization scheme described previously, to keep students engaged and motivated. In summary, for GIFT to function in hybrid human/artificial intelligence instructional teams, an understanding of the human-led portion of the instruction is essential.

Acknowledgements

This research was supported by the US Army Research Laboratory (W911NF-12-20030), the Institute of Education Sciences (R305A130030), and the Office of Naval Research (N00014-12-C-0643). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of these sponsoring agencies.

References

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B. & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system-secondary. *School Psychology Review, 42*(1), 76–98.

American Institutes for Research. (2016). *Databases on state teacher and principal evaluation policies.* Retrieved 2016-12-27, from http://resource.tqsource.org/stateevaldb/Compare50States.aspx.

Archer, J., Cantrell, S., Holtzman, S. L., Joe, J. N., Tocci, C. M. & Wood, J. (2016). *Better feedback for better teaching: A practical guide to improving classroom observations.* Jossey-Bass. Retrieved 2016-12-27, from http://k12education.gatesfoundation.org/wp-content/uploads/2016/05/BetterF.

Bakeman, R., McArthur, D., Quera, V. & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods, 2*(4), 357–370.

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays*. University of Texas Press.

Blanchard, N., D'Mello, S., Olney, A. M. & Nystrand, M. (2015). Automatic classification of question & answer discourse segments from teacher's speech in classrooms. In O. C. Santos et al. (Eds.), *Proceedings of the 8th international conference on educational data mining* (pp. 282–288). International Educational Data Mining Society.

Blanchard, N., Donnelly, P., Olney, A. M., Samei, B., Ward, B., Sun, X., ... D'Mello, S. K. (2016a, September). Identifying teacher questions using automatic speech recognition in classrooms. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue* (pp. 191–201). Los Angeles: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W16-3623.

Blanchard, N., Donnelly, P. J., Olney, A. M., Samei, B., Ward, B., Sun, X., ... D'Mello, S. K. (2016b). Semi-automatic detection of teacher questions from human-transcripts of audio in live classrooms. In *The 9th international conference on educational data mining* (p. 288–291).

Cantrell, S. & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Retrieved 2016-12-27, from http://metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

Caughlan, S., Juzwik, M. M., Borsheim-Black, C., Kelly, S. & Fine, J. G. (2013). English teacher candidates developing dialogically organized instructional practices. *Research in the Teaching of English*, *47*(3), 212–246.

Clotfelter, C. T., Ladd, H. F. & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4), 778–820.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching.* ASCD.

D'Mello, S. K., Olney, A. M., Blanchard, N., Samei, B., Sun, X., Ward, B. & Kelly, S. (2015). Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 557–566). New York, NY, USA: ACM. Retrieved from http://doi.ACM.org/10.1145/2818346.2830602 doi: 10.1145/2818346.2830602.

Donnelly, P. J., Blanchard, N., Olney, A. M., D'Mello, S. K., Nystrand, M. & D'Mello, S. K. (in press). Words matter: Automatic detection of questions in classroom discourse using linguistics, paralinguistics, and context. In *Lak '17: Proceedings of the seventh international conference on learning analytics & knowledge.* ACM.

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., ... D'Mello, S. K. (2016a). Automatic teacher modeling from live classroom audio. In *Proceedings of the 2016 conference on user modeling adaptation and personalization* (pp. 45–53). New York, NY, USA: ACM. Retrieved from http://doi.ACM.org/10.1145/2930238.2930250 doi: 10.1145/2930238.2930250.

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., ... D'Mello, S. K. (2016b). Multi-sensor modeling of teacher instructional segments in live classrooms. In *Proceedings of the 18th ACM international conference on multimodal interaction* (pp. 177–184). New York, NY, USA: ACM. Retrieved from http://doi.ACM.org/10.1145/2993148.2993158 doi: 10.1145/2993148.2993158.

Gamoran, A. (2013). *Educational inequality in the wake of No Child Left Behind.* Spencer Foundation Lecture to the Association for Public Policy and Management, Washington, DC. Retrieved from: http://www.appam.org/awards/spencer-foundation-lectureship.

Gamoran, A. & Kelly, S. (2003). Tracking, instruction, and unequal literacy in secondary school English. In R. Dreeben & M. T. Hallinan (Eds.), *Stability and change in American education: Structure, process, and outcomes* (pp. 109–126). Clinton Corners, NY: Eliot Werner Publications Incorporated.

Gamoran, A. & Nystrand, M. (1991). Background and instructional effects on achievement in eighth-grade English and social studies. *Journal of Research on Adolescence*, *1*(3), 277–300.

Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Bill and Melinda Gates Foundation. January.

Goe, L., Biggers, K. & Croft, A. (2012). *Linking Teacher Evaluation to Professional Development: Focusing on Improving Teaching and Learning*. Research & Policy Brief. National Comprehensive Center for Teacher Quality.

Grossman, P., Loeb, S., Cohen, J. & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*, 445–470.

Hamilton, L. (2012). Measuring teaching quality using student achievement tests: Lessons from educators' response to No Child Left Behind. In *Assessing teacher quality: Understanding teacher effects on instruction and achievement*, edited by S. Kelly (pp. 49–76). New York, NY: Teachers College Press.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, *24*(3), 1141–1177.

Hanushek, E. A. & Rivkin, S. G. (2006). Teacher quality. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 2, pp. 1051–1078). Amsterdam: North Holland.

Hanushek, E. A. & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review, 100*(2), 267–271.

Hardt, M., Price, E. & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315–3323).

Harris, D. N., Ingle, W. K. & Rutledge, S. A. (2014). How Teacher Evaluation Methods Matter for Accountability: A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures. *American Educational Research Journal, 51*(1), 73–112.

Hill, C. J., Bloom, H. S., Black, A. R. & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172–177.

Hill, H. C., Blunk, M. L., Charalambos, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L. & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430–511.

Jennings, J. L. & Corcoran, S. P. (2012). Beyond high stakes tests: Teacher effects on other educational outcomes. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement* (pp. 77–96). New York: Teachers College Press.

Kane, T. J., Kerr, K. A. & Pianta, R. C. (Eds.). (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. Jossey-Bass.

Kane, T. J. & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation. Retrieved 2016-12-27, from http://k12education.gatesfoundation.org/wp-content/uploads/2016/06/MET_-Gathering_Feedback_for_Teaching_Summary1.pdf.

Kane, T. J., McCaffrey, D. F., Miller, T. & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Bill & Melinda Gates Foundation.

Kelly, S. (2012). Understanding teacher effects: Market versus process models of educational improvement. In S. Kelly (Ed.), *Assessing Teacher Quality: Understanding Teacher Effects on Instruction and Achievement* (pp. 7–32). NY: Teachers College Press.

Konstantopolous, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record, 116* (1).

Stein, M. K. & Matsumura, L. C., (2009). Measuring instruction for teacher learning. In D.H. Gitomer (Ed.) *Measurement issues and assessment for teacher quality.* (pp. 179–205). Thousand Oaks: Sage Publications. Retrieved from http://d-scholarship.pitt.edu/26219/.

Mihaly, K., McCaffrey, D., Sass, T. R. & Lockwood, J. R. (2013). Where You Come From or Where You Go? Distinguishing Between School Quality and the Effectiveness of Teacher Preparation Program Graduates. *Education Finance and Policy, 8*(4), 459–493. https://doi.org/10.1162/EDFP_a_00110

Minsky, M. (1961, Jan). Steps toward artificial intelligence. *Proceedings of the IRE, 49*(1), 8-30. doi: 10.1109/JRPROC.1961.287775.

Nye, B., Konstantopoulos, S. & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.

Nystrand, M. (1988). *CLASS (classroom language assessment system) 2.0: A windows laptop computer system for the in-class analysis of classroom discourse*. Wisconsin Center for Education Research.

Nystrand, M. & Gamoran, A. (1997). The big picture: Language and learning in hundreds of English lessons. In M. Nystrand (Ed.), *Opening dialogue: Understanding the dynamics of language and learning in the English classroom* (pp. 30–74). New York: Teachers College Press.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Ryan, K. J. (2016). *Who's smartest: Alexa, Siri, and or Google Now?* Retrieved 2016-12-29, from http://www.inc.com/kevin-j-ryan/internet-trends-7-most-accurate-word-recog.

Sartain, L., Stoelinga, S. R. & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Consortium on Chicago School Research.

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., ... Graesser, A. (2014). Domain independent assessment of dialogic properties of classroom discourse. In J. Stamper, Z. Pardos, M. Mavrikis & B. McLaren (Eds.), *Proceedings of the 7th international conference on educational data mining* (pp. 233–236).

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N. & Graesser, A. (2015). Modeling classroom discourse: Do models that predict dialogic instruction properties generalize across populations? In O. C. Santos et al. (Eds.), *Proceedings of the 8th international conference on educational data mining* (pp. 444–447). International Educational Data Mining Society.

Sinatra, A. M. (2015). A Personalized GIFT: Recommendations for Authoring Personalization in the Generalized Intelligent Framework for Tutoring. In *Foundations of Augmented Cognition* (pp. 675–682). Springer, Cham. https://doi.org/10.1007/978-3-319-20816-9_64.

CHAPTER 24 – Assessment of Collaborative Problem Solving

Arthur C. Graesser[1], Zhiqiang Cai[1], Xiangen Hu[1], Peter W. Foltz[2], Samuel Greiff[3], Bor-Chen Kuo[4], Chen- Huei Liao[4], and David Williamson Shaffer[5]
University of Memphis[1], University of Colorado Boulder and Pearson[2], University of Luxembourg[3], National Taichung University of Education[4], University of Wisconsin[5]

Introduction

Collaborative problem solving (CPS) is one of the important 21st century skills that has attracted interest in international assessments, national assessments of middle and high school students, colleges, business, and the military (Griffin & Care, 2015; Hesse, Care, Buder, Sassenberg & Griffin, 2015; NRC, 2011; OECD, 2013; Sottilare et al., 2015). CPS is an essential skill in the home, the workforce, and the community because much of the planning, problem solving, and decision making in the modern world is performed by teams. The success of a team can be threatened by a social loafer, an uncooperative unskilled member, or a counterproductive alliance, whereas it can be facilitated by a strong team member that draws out different perspectives, helps negotiate conflicts, assigns roles, promotes team communication, and guides the team to overcome troublesome obstacles (Fiore, Wiltshire, Oglesby, O'Keefe & Salas, 2014; Salas, Cooke & Rosen, 2008).

CPS differs from individual problem solving (IPS) in ways that may have both positive and negative consequences. CPS allegedly has advantages over IPS because 1) there is a more effective division of labor, 2) the solutions incorporate information from multiple sources of knowledge, perspectives, and experiences, and 3) the quality of solutions is stimulated by ideas of other team members. There are also potential disadvantages of CPS to the extent that 1) team members waste time with irrelevant discussion, 2) there is diffusion of responsibility in completing tasks, and 3) disagreements among team members occur that paralyze progress in solving the problem.

At the international level, CPS was selected by the Organisation for Economic Co-operation and Development (OECD) as a new development for the Program for International Student Assessment (PISA) in the 2015 international survey of student skills and knowledge (Graesser, Forsyth & Foltz, 2016; OECD, 2013, 2015). Fifteen-year-old students from over three dozen countries completed this PISA CPS 2015 assessment in addition to assessments of mathematics, science, literacy, and other proficiencies. One of the goals of this chapter is to describe how CPS was assessed in PISA CPS 2015.

PISA used computer agents in the 2015 assessment. That is, a single human interacts with one, two, or three computer agents as team members rather than other humans. Conversation-based assessments with computer agents are manifested by chat conversations as well as actions of team members (Zapata-Rivera, Jackson & Katz, 2015). Computer agents are believed to provide control over the social interaction so that important assessments can be made with consistency and control, two requirements that communicating with fellow humans could not provide. Agents also provide control over logistical and measurement problems that stem from 1) assembling groups of humans (via computer mediated conversation) in a timely manner, 2) the necessity of having multiple teams per student to obtain reliable assessments in different circumstances, and 3) extreme measurement error when particular students are paired with other humans who do not collaborate well. A second goal of this chapter is to describe how agents can be used to provide meaningful assessments of CPS.

Although conversation-based assessments with agents can provide meaningful assessments of CPS, there is still an important goal of assessing interactions among humans. That requires an automated analysis of natural language and discourse in addition to identifying how particular problem-solving patterns map onto

important CPS proficiencies (e.g., establishing shared knowledge, taking initiative, communicating important information to the group). The third goal of this chapter is to identify some of the automated approaches that show promise in automated assessments of CPS among humans. These methods could be integrated with the Generalized Intelligent Framework for Tutoring (GIFT) in future developments by piggybacking on and expanding existing applications of natural language processing in GIFT.

Related Research

There have been a number of theoretical frameworks for analyzing CPS. Some of the prominent ones include the Center for Research on Evaluation, Standards, and Student Testing's (CRESST) teamwork processing model (O'Neil, Chuang & Baker, 2010), the teamwork models of Salas, Fiore, and colleagues (Fiore et al., 2010; Salas, Cooke & Rosen, 2008) and the Assessment and Teaching of 21st Century Skills (ATC21S; Griffin & Care, 2015; Hesse et al., 2015). All of these frameworks have both a *cognitive* dimension that includes problem solving and other cognitive processes and a *collaborative* dimension that includes communication and other social interaction processes. These approaches were incorporated in PISA CPS 2015 (Graesser et al., 2016; OECD), the framework under direct focus in this chapter.

The problem-solving dimension in PISA CPS 2015 framework incorporated the same PISA 2012 problem solving framework that targeted individual problem solving (Funke, 2010; OECD, 2010; Greiff, Kretzschmar, Müller, Spinath & Martin, 2014). There were four cognitive processes (or competencies) on the problem-solving dimension:

1) **Exploring and understanding.** Interpreting the initial information about the problem and any information that is uncovered during the course of exploring and interacting with the problem.
2) **Representing and formulating**. Identifying global approaches to solving the problem, relevant strategies and procedures, and relevant artefacts (e.g., graphs, tables, formulae, symbolic representations) to assist in solving the problem.
3) **Planning and executing.** Constructing and enacting goal structures, plans, steps, and actions to solve the problem. The actions can be physical, social, or verbal.
4) **Monitoring and reflecting**. Tracking the steps in the plan to reach the goal states, marking progress, and reflecting on the quality of the progress or solutions.

There were three processes on the collaborative dimension:

1) **Establishing and maintaining shared understanding**. Keeping track of what each other knows about the problem (i.e., shared knowledge, common ground; Clark, 1996), the perspectives of team members, and a shared vision of the problem states and activities (Cannon-Bowers & Salas, 2001; Dillenbourg & Traum, 2006).
2) **Taking appropriate actions to solve the problem.** Performing actions that follow the appropriate steps to achieve a solution. This includes physical actions and communication acts that advance the solution to the problem.
3) **Establishing and maintaining group organization.** Helping organize the group to solve the problem by considering the talents and resources of group members during the assignment of roles; following the rules of engagement for one's own roles as well as handing obstacles to tasks assigned to other team members.

When the 4 problem-solving processes are crossed with the 3 collaboration processes, there are 12 skills in the resulting *CPS assessment matrix*. Table 1 shows this matrix that was adopted in the PISA CPS 2015 framework. A satisfactory assessment of CPS would assess the skill levels of students for each of these 12 cells and these would contribute to a student's overall *CPS proficiency measure*.

Table 1. Copied from OECD (2013). PISA 2015 collaborative problem solving framework.

	(1) Establishing and maintaining shared understanding	**(2) Taking appropriate action to solve the problem**	**(3) Establishing and maintaining team organization**
(A) Exploring and Understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve problem
(B) Representing and Formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describe roles and team organization (communication protocol/rules of engagement)
(C) Planning and Executing	(C1) Communicating with team members about the actions to be/ being performed	(C2) Enacting plans	(C3) Following rules of engagement, (*e.g.*, prompting other team members to perform their tasks.)
(D) Monitoring and Reflecting	(D1) Monitoring and repairing the shared understanding	(D2) Monitoring results of actions and evaluating success in solving the problem	(D3) Monitoring, providing feedback and adapting the team organization and roles

As mentioned, the PISA CPS 2015 assessment had students interact with computer agents rather than other humans. The following definition of CPS was articulated in the PISA CPS 2015 framework (OECD, 2013: *Collaborative problem-solving competency is the capacity of an individual to effectively engage in a process whereby **two or more agents** attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution.* An agent could be either a human team member or a computer agent that interacts with the student. The final assessment that was adopted had students interact with one to three computer agents instead of other humans. Therefore, the overall CPS proficiency measure assessed how well a single human interacted with computer agents during the course of problem solving. The computer agents were minimalist agents in a chat facility, without text-to-speech, animation, or visual depictions of what they looked like. This was necessary to eliminate biases of culture, personality, and emotions, which were beyond the scope of the PISA CPS assessment.

A central advantage of assessments with computer agents is the degree of control over the conversation. The discourse contributions of the two agents (A1, A2) and any associated digital media (M) can be coordinated so that each [A1, A2, M] sequential display is functionally a single episodic unit (U) to which the human responds through language, action, or silence in a particular human turn (HT). Thus, there is an orchestrated finite-state transition network that alternates between episodic units (U) and human turns (HT), which is formally isomorphic to a dialogue. This is very different than a collaboration in which many people can speak simultaneously and overlap in time (Clark, 1996). Conditional branching can occur in the state-transition network (STN) so that the computer's generation of U_{n+1} at turn n+1 is contingent on the state of the human turn HT_n at turn n. However, the degree of branching was limited to a small number of states

associated with each human turn (HT_n) in PISA CPS 2015; there were 2–4 alternative multiple-choice options at each turn (i.e., either chat options or alternative actions to be performed). Consequently, the fan out of conditional branching was not complex and the turn-taking frequently converged at points of assessment rather than diverging in many directions. Only one score was associated with each episodic unit and each episodic unit was aligned with one and only one of the 12 cells in the CPS assessment matrix.

The design of the PISA CPS 2015 assessment was compatible with the normal psychometric modeling in the world of assessment, where multiple-choice tests are ubiquitous. Traditional psychometric assessments routinely include a fixed set of items (i.e., episodic units) that all humans experience. Analogously, PISA CPS 2015 had a fixed sequence episodic units (U_1, U_2, … U_m) that occurred at specific points as the problem was solved and the responses of the human were automatically recorded (as clicks on action options or chat options). The conversations were designed so that the conversations would naturally close shortly after the human responded to an episodic unit and the subsequent episodic unit was launched (e.g., "Thanks for your input, let's go on"). Assessment scores were collected for each student for the M episodic units that collectively covered each of the 12 cells in the CPS assessment matrix. These scores contributed to overall CPS proficiency measures that have not yet been finalized by OECD.

Students encounter a diverse set of situations in PISA CPS 2015 in order to make sure that important conditions are covered in the assessment. Students who respond randomly to the response options would obviously receive low values on CPS proficiency as well as the collaboration and problem solving dimensions. A student may be a good team player and be responsive, but not take the initiative when there are problems (e.g., an agent who is unresponsive, or a new obstacle in the problem occurs). A student may take some initiative when there are breakdowns, but not be able to handle very complex cognitive problems. A student who scores high in CPS proficiency takes the initiative in moving the team to achieve group goals during difficult times (conflicts, incorrect actions, unresponsive team members) and can also handle complex problems with many cognitive components that burden working memory and require reasoning. Episodic units for all of these situations are needed in order to have an adequate CPS assessment. In contrast, many of these situations might not arise when a student interacts with other humans so there would be missing scores for some of the 12 cells.

Computer agents may be suitable for providing a summative assessment of CPS proficiency that is both reliable and valid. Available data have so far supported the validity of the PISA CPS 2015 framework. For example, a factor analysis has shown an extremely high correspondence between a human-agent CPS assessment and a human-human assessment in a sample in Germany students (Greiff, personal communication). Kuo et al. (2016) conducted an assessment in Taiwan that adopted the PISA CPS 2015 assessment framework. The study developed an internet-based CPS assessment with conversational agents on five tasks to be completed in 100 minutes. There were over 50,000 ninth and tenth grade students who participated between October 2014 and February 2015. The problem-solving dimension in the PISA CPS 2015 assessment showed a similar ordering of competencies for the four problem-solving components (A > B > C > D) as were found for the PISA 2012 assessments of individual problem solving. Although the complete data for PISA CPS 2015 is still being analyzed for over 400,000 students in three to four dozen different countries, the reliability of the data in field trials is encouraging.

Discussion

Although computer agents may be suitable for a summative assessment of CPS proficiency, there are major limitations with this approach for teams of humans and formative CPS assessment. Computer-based environments for teams (whether they be collaborative learning, problem solving, or work) need automatic tracking and analysis of the language, actions, and social interactions of human team members. Computer-based environments need to adaptively, intelligently, and immediately respond to the team members based

on the automated assessments of CPS proficiencies and many other cognitive and noncognitive characteristics of team members. The agent-based assessment in PISA CPS 2015 does not offer any help in developing a computer environment for tracking and responding to teams of humans. The latter would be needed in GIFT (Sottilare et al, 2017).

The remainder of this chapter identifies some promising ways of automatically tracking the language and discourse of humans in team chat interactions. Ideally, we would be able to map particular language and conversation patterns onto the cells of CPS assessment matrix. If these patterns could be detected automatically, then there is a principled theoretical foundation for 1) a formative assessment of CPS skills of team members and 2) recommendations on how the computer environment should respond to unproductive teams or team members.

A community of researchers in the learning sciences and computational linguistics have investigated conversations in small groups by analyzing the log files of computer-mediated interactions in chat and discussion forums (Dowell, Graesser & Cai, in press; Foltz & Martin, 2008; Liu, Von Davier, Hao, Kyllonen & Zapata-Rivera, 2015; Mu, Stegmann, Mayfield, Rosé & Fischer, 2012; Shaffer, Collier & Ruis, in press; Tausczik & Pennebaker, 2013; Von Davier & Halpin, 2015). The conversations have been analyzed by a variety of automated text analysis tools, such as state-transition networks that track speech acts of team players (Morgan, Keshtkar, Duan & Graesser, 2012), latent semantic analysis (Foltz & Martin, 2008; Gorman et al., 2003), epistemic network analysis (Shaffer et al., 2009), Coh-Metrix (Graesser, McNamara, et al., 2014), and Linguistic Inquiry and Word Count (Pennebaker, Booth & Francis, 2007). These automated tools have been applied to conversations in their entirety, to subsets of the conversation at a particular window size (e.g., 5 turns in a row), to single conversational turns, to adjacent conversational turns, and to turns of specific team members. The conversation profile includes measures of team cohesion, percentage of on-topic versus off-topic contributions, amount of new information, characteristics of team members (e.g., driver, follower, social loafer), alliances between team members, and presence of specific conversation patterns. It is beyond the scope of this chapter to describe in detail these automated approaches (see Graesser, Dowell, Clewley & Shaffer, submitted), but we do highlight some examples to illustrate the prospects of this approach.

Matches to Expectations

In many applications of team problem solving, there are a set of *expectations* that need to be covered to solve the problem. An expectation is a sentence, clause, proposition, or expression of comparable length, as discussed in reports on AutoTutor (Graesser, 2016; Cai, Graesser & Hu, 2015). A solution to a problem consists of a set of expectations that hopefully would be covered by the team. The team or team member received higher scores to the extent that more expectations are articulated during the chat conversation. Physical actions are also handled in this way: performance increases as more critical actions are performed.

Advances in computational linguistics and semantics have made impressive gains in the accuracy of semantic matches between one short text (i.e., a sentence or two) and another short text (Rus, Lintean, Graesser & McNamara, 2012; Rus & Ştefănescu, 2016). The accuracy is not always perfect, but it often is impressive and on par with human experts who judge the semantic similarity of pairs of short texts. The AutoTutor research team has evaluated many semantic matchers over the years in AutoTutor and other intelligent tutoring systems (ITSs) with conversational agents (Cai et al., 2011; Graesser, Penumatsa, Ventura, Cai & Hu, 2007; Rus et al., 2012). The semantic matchers automatically compute the semantic similarity between a student's verbal contribution and an expectation, with a similarity score that varies from zero to one. These semantic match algorithms have included keyword overlap scores, word overlap scores that place higher weight on lower frequency words in the English language, scores that consider the order of words, latent semantic analysis cosine values, comparisons to regular expressions, and procedures that

compute semantic logical entailment. As an example, Cai et al. (2011) reported that the correlation of similarity scores between AutoTutor and human expert judges was $r = 0.667$, about the same as between two trained judges ($r = 0.686$). Interestingly syntactic parsers did not prove useful in these analyses because a high percentage of the students' contributions are vague, telegraphic, elliptical, and ungrammatical. At the time of this writing, the best automated semantic matcher is the Semantic Similarity (SEMILAR) system developed by Rus et al. (2013). SEMILAR won the semantic textual similarity competition at *SemEval-2015*, the premier international forum for semantic evaluation.

Matches to expectations are powerful in assessments of CPS to the extent that the solutions to a problem are known ahead of time, as in the case of PISA CPS 2015. Indeed, there could be a set of expectations associated with each of the 12 cells in the CPS assessment matrix and these could be scored for each team member over the course of the CPS interaction. Unfortunately, this approach does not work when there are no expectations in a CPS application. The subsequent approaches can be applied when a problem does not have a finite set of associated expectations.

Automated Speech Act Classification and State Transition Networks

The content of each turn is classified into speech acts and each speech act is assigned to a category (Liu et al., 2015; Morgan et al., 2012). For example, the speech act categories defined by Rus, Graesser, Moldovan, and Niraula (2012) were Statement, Question, Request, Reaction, MetaStatement, Expressive Evaluation, and Greeting. Automated speech act classification has achieved a moderate degree of accuracy compared with trained human annotators (Olney et al., 2003; Rus et al., 2012; Samei et al., 2014). A chat window of five turns appears to be an optimal chat length to analyze the context of particular turns in computer-mediated chat during collaborative learning and CPS (Collier, Ruis & Shaffer, 2016; Samei et al., 2014). This amount of context has been explored to improve speech act classification accuracy and to detect multi-turn discourse patterns. Another approach is to construct a STN on adjacent speech acts (Morgan et al., 2012). An STN computes the probabilities of adjacency pairs in a corpus of chat sequences. Stated more formally, it is the transition probability between adjacent speech act categories (SAC) that are indexed by particular team participants: [P-SAC_n → P-SAC_{n+1}]?

Some measures of CPS can theoretically be derived from the distribution P-SAC node categories and the transition probabilities between these node categories. Students who take initiative would have a high proportion of Questions, Requests and Statements, whereas students who are responsive team members (but not leaders) would have a relatively high proportion of Reactions. A disruptive team member would have a high proportion of negative Expressive Evaluations, whereas a social loafer would have a low number of contributions compared with other team members. Regarding the state transitions, responsive team members would have a relatively high transition probability between Questions/Requests of others and the participant's Reactions or Statements; these transition probabilities would be low for unresponsive team members. Thus, these probabilistic metrics have relevance to many of the cells in the CPS assessment matrix. However, available studies have not empirically evaluated the mapping between these automated measures and the 12 cells in the CPS assessment matrix.

Latent Semantic Analyses (LSA) and Semantic Comparisons

LSA (Landauer, Foltz & Laham, 1998) is used to analyze the semantic content of the team members' contributions, a level of language that is not tapped in speech act analyzes. For example, LSA has been used to analyze the coherence of teams and characteristics of individual team members (Dowell, 2017; Foltz & Martin, 2008; Gorman et al., 2003). Text excerpts can be evaluated on semantic similarity through LSA as well as other semantic similarity evaluators that have been described.

Semantic comparison metrics, such as LSA and SEMILAR (Rus et al., 2013), provide an assessment of establishing a shared understanding and building on what each other knows, both of which are theoretically important in the process of establishing and maintaining shared understanding component of collaboration in PISA CPS 2015. The *relevance* (R) of a turn's meaning to the problem being solved is used to compute the extent to which a turn is on versus off topic. This is measured as the semantic overlap between each turn and the semantic topics in the problem being solved. The *givenness* (G) and *newness* (N) of individual turns can be computed for individual team members and the team as a whole (Hempelmann et al., 2005; Hu et al., 2014). A productive collaborative team member contributes relevant information that is new and also builds on other team member's topic-relevant ideas in a responsive fashion. Scores for R, G, and N can be automatically computed by LSA and other semantic evaluators such as SEMILAR, with values that vary from near zero to one. For example, a team member who productively leads the conversation would have a vector of RGN measures such as (0.9, 0.4, 0.6). Team members who echo ideas of others in a conversation would have a (0.9, 0.5, 0.0) vector if they stay on topic, but a (0.0, 0.5, 0.0) vector on off-topic talk. A team member with a (0.0, 0.0, 0.9) vector would be in their own irrelevant worlds and not helpful to collaboration. These profiles have been confirmed in a recent dissertation by Dowell (2017).

There are many other measures of team members and teams that can be computed from these similarity-based metrics and transitions between team members (Dowell, 2017). *Participation* is the relative proportion of a participant's contributions (turns) out of the total number of group contributions; physical actions can be computed in an analogous way to assess the second component of the collaborative dimension (taking appropriate actions to solve the problem) in PISA CPS 2015. *Responsiveness* (analogous to G for givenness) assesses how responsive a team member's contributions are to all other group members' previous contributions in the conversation. *Social impact* measures how turn contributions of a team member have a semantic similarity to other members' contributions in future follow-up responses. *Team member cohesion* measures how semantically similar a team member's contributions are to the same member's previous conversational turns. That is, is a team member saying the same thing over and over? *Communication density* measures how much information in a turn is distinctive to the topic, compared with everyday topics of conversation. All of these metrics can be applied to individual team members as well as the team as a whole.

Epistemic Network Analysis (ENA)

ENA attempts to assess the complex thinking, discourse, reasoning, and topics addressed in professional disciplines and communities (Nash & Shaffer, 2011; Shaffer et al., 2009; Shaffer, Collier & Ruis, in press). There is a disciplinary style of thinking and talking that resonates with the expertise of the community of stakeholders. In scientific disciplines, for example, the discourse might involve claims with supporting empirical evidence and causal analyses. That is a very different discourse from mathematicians or art historians.

ENA's analysis of chat in CPS begins by representing the content as a network structure of connections among critical knowledge, skills, values, and epistemic moves in a professional domain. It measures the strength of association among these cognitive elements and quantifies changes in the composition and strength of those connections over time for individual team members and the entire team. ENA constructs a metric space that enables comparison of individual or group networks through 1) difference graphs, which visualize the differences in weighted connections between two networks, and 2) summary statistics, which reflect the weighted structure of connections in the networks.

It is beyond the scope of this chapter to precisely specify the algorithms that underlie ENA and the process of applying ENA to CPS data (Shaffer, Collier & Ruis, in press, for the ENA Toolkit). The initial step

consists of annotating chat turn sequences (i.e., stanzas, sliding turn windows of about length 5) on important cognitive categories (i.e., expressions of skills, knowledge, identity, values, and epistemic content), based on the words expressed in those turns. The next step computes a matrix of co-occurrences of these cognitive categories within these turn sequences and statistically reduces the resulting set of co-occurrence matrices to a small number of dimensions through singular value decomposition. When there are only two or three dimensions, it is possible to plot each cognitive category in a 2- or 3-D metric space; the size of the cognitive category in the space reflects its relative frequency, whereas the thickness of the links between the concept categories reflects the co-occurrence frequency. The resulting network patterns can be compared for different team members, the team as a whole, different phases of CPS interactions, and different chat contexts associated with the profession. ENA has been applied to the land science chat corpora (Collier, Ruis & Shaffer, 2016) and medical engineering design in teams of 3–5 members along with a mentor.

A discipline-oriented style of thinking and talking would of course be an important characteristic to detect and track in team-based ITSs integrated with GIFT. A team member or team as a whole would be regarded as having higher domain expertise to the extent that the chat exhibits higher disciplinary talk that can be automatically quantified from the qualitative input. There are also some links to the CPS assessment matrix of PISA CPS 2015. For example, the discipline thinking parameters are relevant to the problem-solving component D (monitoring and reflecting) and the identity concept categories are relevant to collaboration component 3 (establishing and maintaining team organization). At this point there has been no systematic evaluation of the use of ENA in the scoring of CPS based on the PISA 2015 framework.

Recommendations and Future Research

The most obvious recommendation is to add these automated measures of CPS to GIFT and team-based ITSs. The scores, competencies, and measurements at varying grain sizes would be automatically computed and stored in the Learner Record Store of GIFT. This chapter identifies some automated quantitative algorithms for detecting and assessing many aspects of CPS from the content of the chat logs. These assessments apply to either the entire team or to individual team members as units of analysis. These content-based assessments are a generation beyond the traditional sociometric analyses that compute simple metrics, such as who talks most, who talks to whom, and how many words. Most of the assessments are also aligned with the theoretical framework of a large-scale international assessment, namely PISA CPS 2015. Although this is a promising start, the reliability and validity of these automated assessments await future research.

A second major recommendation is to incorporate these CPS assessments in production rules that formulate what the adaptive, intelligent tutor does next. For example, the suite of applications in GIFT already has the AutoTutor agents that help individuals learn by holding a conversation in natural language (Graesser, 2016; Cai et al., 2015). This could be expanded to include an automated AutoMentor in team contexts when GIFT takes on teams. Some simple production rules were proposed in Graesser et al. (submitted):

1) If the team is stuck and not producing contributions on the relevant topic, then the agent says, "What's the goal here?" or "Let's get back on track."
2) If the team meanders from topic to topic without much coherence, then the agent says, "I'm lost!" or "What are we doing now?"
3) If the team is saying pretty much the same thing over and over, then the agent says, "So what's new?" or "Can we move on?"
4) If a particular team member (Harry) is loafing, the agent says, "What do you think, Harry?"
5) If a particular team member (Sally) is dominating the conversation excessively, the agent says, "I wonder what other people think about this?"
6) If one or more team members express unprofessional language, the agent says, "Let's get serious now. I don't have all day."

Important next steps are to identify a larger set of production rules for CPS, implement them in GIFT environments, and evaluate whether they improve collaborative problem-solving performance (Sottilare et al., 2017).

Acknowledgements

The research on was supported by the National Science Foundation (NSF) (DRK–12-0918409, DRK–12 1418288), the Institute of Education Sciences (IES) (R305C120001), the US Army Research Laboratory (W911INF-12-2-0030), and the Office of Naval Research (N00014-12-C-0643; N00014-16-C-3027). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or the US Department of Defense. The Tutoring Research Group (TRG) is an interdisciplinary research team comprised of researchers from psychology, computer science, and other departments at University of Memphis (visit http://www.autotutor.org).

References

Cai, Z., Graesser, A.C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., Halpern, D. & Butler, H. (2011, November). Trialog in ARIES: user input assessment in an intelligent tutoring system. In W. Chen & S. Li (Eds.), *Proceedings of the 3rd IEEE International Conference on Intelligent Computing and Intelligent Systems* (pp.429–433). Guangzhou: IEEE Press.

Cai, Z., Graesser, A.C. & Hu, X. (2015). ASAT: AutoTutor script authoring tool. In. R. Sottilare, A.C. Graesser, X. Hu & K. Brawner (Eds.), Design Recommendations for Intelligent Tutoring Systems: Authoring Tools (Vol. 3) (pp.199–210). Orlando, FL: US Army Research Laboratory.

Cannon-Bowers, J. A. & Salas, E. (2001). Reflections on shared cognition. *Journal of Organizational Behavior, 22*, 195–202.

Clark, H.H. (1996). *Using language*. Cambridge, United Kingdom: Cambridge University Press.

Collier, W., Ruis, A. & Shaffer, D. W. (2016). Local versus global connection making in discourse. In International Conference of the Learning Sciences. Singapore.

Dillenbourg, P. & Traum, D. (2006). Sharing solutions: Persistence and grounding in multi-modal collaborative problem solving. *The Journal of the Learning Sciences, 15*, 121–151.

Dowell, N.M. (2017). A computational linguistics analysis of learners' discourse in computer-mediated group learning environments. Dissertation, University of Memphis.

Dowell, N. M., Graesser, A. C., Cai, Z. (in press). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. Journal of Learning Analytics.

Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M. & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors, 52*, 203–224.

Fiore, S. M., Wiltshire, T. J., Oglesby, J. M., O'Keefe, W. S. & Salas, E. (2014). Complex collaborative problem-solving processes in mission control. *Aviation, space, and environmental medicine, 85*(4), 456–461.

Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. Discourse Processes, 25, 285–307.

Foltz, P. W. & Martin, M. J. (2008). Automated communication analysis of teams. In E. Salas, G. F. Goodwin & S. Burke (Eds.), Team effectiveness in complex organisations and systems: Cross-disciplinary perspectives and approaches (pp. 411–431). New York, NY: Routledge.

Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processes, 11*, 133–142.

Gorman, J. C., Foltz, P. W. Kiekel, P. A., Martin, M. A. & Cooke, N. J. (2003) Evaluation of Latent Semantic Analysis-based measures of communications content. *In Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting.*

Graesser, A.C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education, 26*, 124–132.

Graesser, A.C., Dowell, N.M., Clewley, D. & Shaffer, D.W. (submitted). Agents in collaborative problem solving. *International Journal of Computer Supported Collaborative Learning.*

Graesser, A.C., Forsyth, C.M. & Foltz, P. (2016). Assessing conversation quality, reasoning, and problem solving performance with computer agents. In B. Csapo, J. Funke, and A. Schleicher (Eds.), *On the nature of problem solving: A look behind PISA 2012 problem solving assessment* (pp. 275–297). Heidelberg, Germany: OECD Series.

Graesser, A.C., McNamara, D.S., Cai, Z., Conley, M., Li, H. & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. Elementary School Journal, 115, 210–229.

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z. & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed initiative dialogue in natural language. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), Handbook of latent semantic analysis (pp. 243–262). Mahwah, NJ: Erlbaum.

Greiff, S., Kretzschmar, A., Müller, J. C., Spinath, B. & Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *Journal of Educational Psychology*, *106*(3), 666–680.

Griffin, P. & Care, E. (2015). ATC21S method. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills: Methods and Approach*. Dordrecht: Springer.

Hempelmann, C. F., Dufty, D., McCarthy, P., Graesser, A. C., Cai, Z. & McNamara, D. S. (2005). Using LSA to automatically identify givenness and newness of noun-phrases in written discourse. In B. Bara (Ed.), Proceedings of the 27th Annual Meeting of the Cognitive Science Society (pp. 941–946). Mahwah, NJ: Erlbaum.

Hesse, F., Care, E., Buder, J., Sassenberg, K. & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin and E. Care (eds.), Assessment and teaching of 21st century sills (pp. 37–55). Heidelberg, GA: Springer.

Hu, X., Nye, B. D., Gao, C., Huang, X., Xie, J. & Shubeck, K. (2014). Semantic representation analysis: A general framework for individualized, domain-specific and context-sensitive semantic processing. In D.D. Schmorrow and C.M. Fidopiastis (eds.), *Foundations of augmented cognition: Advancing human performance and decision-making through adaptive systems* (pp. 35–46). Springer International Publishing.

Kuo, B.-C, Liao, C.-H., Pai K.-C., Shih S.-C., Li C.-H. & Mok, M. C. M. (2016). Computer-based collaborative problem solving assessment in Taiwan. Manuscript submitted for publication.

Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. Discourse Processes, 25, 259–284.

Liu, L., Von Davier, A., Hao, J., Kyllonen, P. & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. In R. Yigal, S. Ferrara & M. Mosharraf (Eds.), Handbook of research on technology tools for real-world skill development (pp. 344–359). Hershey, PA: IGI Global.

Morgan, B., Keshtkar, F., Duan, Y. & Graesser, A.C. (2012). Using state transition networks to analyze multi-party conversations in a serious game. In. S. A. Cerri & B. Clancey (Eds.), Proceedings of the 11th International Conference on Intelligent Tutoring Systems (ITS 2012) (pp. 162–167). Berlin: Springer-Verlag.

Mu, J., Stegmann, K., Mayfield, E., Rosé, C. & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning, 7(*2), 285–305.

Nash, P. & Shaffer, D.W.(2013). Epistemic trajectories: Mentoring in a game design practicum. *Instructional Science 41*(4): 745–771

National Research Council (2011). Assessing 21st century skills. Washington, DC: National Academies Press.

OECD (2013). PISA 2015 collaborative problem solving framework. Paris: France: OECD.

OECD (2015). *PISA 2015 released field trial cognitive items.* Paris, France: OECD.

Olney, A., Louwerse, M., Mathews, E., Marineau, J., Hite-Mitchell, H. & Graesser, A. (2003). Utterance classification in AutoTutor. In J. Burstein & C. Leacock (Eds.), Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing. Philadelphia: Association for Computational Linguistics.

O'Neil, H. F., Chuang, S.H. and Baker, E.L. (2010). Computer-based feedback for computer-based collaborative problem-solving. In D. Ifenthaler, P. Pirnay-Dummer, N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 261–279). New York, NY: Springer-Verlag.

Pennebaker, J. W., Booth, R. & Francis, M. (2007). LIWC2007: Linguistic inquiry and word count. Austin, Texas: liwc. net. 2007, Austin, TX. Retrieved from liwc. Net

Rus, V., Lintean, M. C., Banjade, R., Niraula, N. B. & Stefanescu, D. (2013). SEMILAR: The Semantic Similarity Toolkit. In ACL (Conference System Demonstrations) (pp. 163–168).

Rus, V., Lintean, M., Graesser, A. C. & McNamara, D.S. (2012). Text-to-text similarity of statements. In P. McCarthy and C. Boonthum-Denecke (Eds.), Applied natural language processing: Identification, investigation, and resolution (pp. 110–121). Hershey, PA: IGI Global.

Rus, V., Moldovan, Graesser, A.C. & Niraula, N. (2012). Automatic discovery of speech act categories in educational games. In K.Yacef,O. Zaïane, H. Hershkovitz, M. Yudelson, and J. Stamper (Eds.) Proceedings of the 5th International Conference on Educational Data Mining (pp. 25–32). Chania, Greece: International Educational Data Mining Society.

Salas, E., Cooke, N. J. & Rosen, M. A. (2008). On teams, teamwork, and team performance: Discoveries and developments. Human Factors: The Journal of the Human Factors and Ergonomics Society, 50(3), 540–547.

Samei, B., Li, H., Keshtkar, F., Rus, V. & Graesser, A. C. (2014). Context-based speech act classification in intelligent tutoring systems. In S. Trausan-Matu, K. Boyer, M. Crosby & K. Panou (Eds.), Proceedings of the Twelfth International Conference on Intelligent Tutoring Systems (pp. 236–241). Berlin: Springer.

Shaffer, D. W., Collier, W. & Ruis, A. R. (in press). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics.*

Sottilare, R.A., Burke, C.S., Salas, E., Sinatra, A.M., Johnston, J.H. & Gilbert, S.B. (2017). Towards a Design Process for Adaptive Instruction of Teams: A Meta-Analysis. International Journal of Artificial Intelligence in Education. DOI: 10.1007/s40593-017-0146-z.

Shaffer, D. W., Hatfield, D., Svarovsky, G., Nash, P., Nulty, A., Bagley, E. A., Frank, K., Rupp, A.A., Mislevy, R. J. (2009). Epistemic Network Analysis: A prototype for 21st century assessment of learning. The International Journal of Learning and Media, 1(1), 1–21.

Tausczik, Y. R. & Pennebaker, J. W. (2013). Improving teamwork using real-time language feedback. *Proceedings of Human Factors in Computing Systems (*CHI), 459–468.

Von Davier, A. & Halpin, P. (2013). Collaborative problem solving and the assessment of cognitive skills: psychometric considerations (Research Report No. ETS RR-13-41) (pp. 1–42). Educational Testing Service.

Zapata-Rivera, D., Jackson, G.T. & Katz, I. (2015). Authoring conversation-based assessment scenarios. In R. Sottilare, X. Hu, A. Graesser & K. Brawner (Eds.), Design Recommendations for Adaptive Intelligent Tutoring Systems. (Vol.3)(pp.169–178). Orlando, FL: US Army Research Laboratory.

CHAPTER 25 – Challenges for Assessing and Tutoring Collective Skills

Jeanine Ayers[1], Martin L. Bink[2], and Frederick J. Diedrich[3]
Sophia Speira, LLC[1], US Army Research Institute for the Behavioral and Social Sciences[2],
Independent Consultant[3]

Introduction

Computer-based intelligent tutoring systems (ITSs) rely on assessment of learner behavior, and indeed, ITSs cannot accurately function without an "understanding" of the specific learner. As a result, it is of critical importance to understand what can and cannot be assessed automatically, given various learning environments and learning objectives, as this will impact ITS functionality. Here, we define an assessment as the result of applying a formula to data (a measurement), and then comparing the score on the measurement to an expert behavior (the assessment). Assessment is especially challenging in environments focused on collective tasks (i.e., those that require organized team or unit performance for accomplishment; Department of the Army, 2012), in which many of the key behaviors of interest involve coordination and communication between human and/or synthetic actors (broadly referred to as collective skills here).

Assessment in ITSs is primarily intended to capture the learner state in comparison to a desired state or an expert state. The learner state is often defined by an individual's efficacy on relevant parameters (e.g., domain knowledge, skills or abilities, etc.). The assessment of the learner state results in a model of the learner and helps to drive adaptive support in the ITS. As assessments improve, more accurate learner models are possible (Sottilare, 2013).

Defining the learner model and providing adaptive support is especially complicated when ITSs are applied to collective skills (Sottilare, Holden, Brawner & Goldberg, 2011). Not only are there technical challenges for creating a team ITS architecture (Gilbert, Winer, Holub, Richardson, Dorneich & Hoffman, 2015), but there are also challenges in assessing team performance to understand a collective "learner" state. Shared mental models may be one way to capture the learner model (Sottilare, Brawner, Goldberg & Holden, 2012), but assessing and using shared mental models still presents challenges for ITSs (Fletcher & Sottilare, 2013; Fletcher & Sottilare, 2017).

Given the challenges associated with collective tasks and the critical role for assessment in supporting ITSs, this chapter presents lessons learned from an effort to develop automated assessments for Army Aviation collective task performance in attack-reconnaissance missions as conducted in a networked virtual environment. The findings illustrate gaps in assessment capabilities that will limit ITS functionality with respect to collective versus individual skills. Accordingly, these gaps suggest limits to role of ITSs and inform how ITSs might best support human instruction for collective skills, resulting in recommendations for future research.

Related Research

Issues of collective performance assessment are not unique to ITSs. Assessment is a key part of understanding collaboration (Hmelo-Silver, Chernobilsky & Jordan, 2008), collaborative learning (Strijbos, 2011), teamwork (Salas, Sims & Burke, 2005), and simulation-based training (Siebert, Diedrich, Stewart, Bink & Zeidman, 2011). There are two issues in the assessment of teams: the level of assessment and the focus of assessment. As defined here, the level of assessment refers to the person(s) being assessed whereas the focus of assessment refers to the construct being assessed.

The level of assessment is either the individual or the team. Even though it is generally accepted that aggregated individual performance is not equivalent to collective performance, it may be necessary to measure, assess, and model individual inputs in team ITSs (Fletcher & Sottilare, 2013). Measuring the individual learner states in team ITSs may be one way of understanding shared mental models. It may be the goal of some team ITSs to increase the efficacy of the individual in team contexts. If that were the case, then measuring and assessing individual performance and providing individual feedback would be necessary. By contrast, if the goal of team ITSs is to increase the efficacy of collective performance, then assessments of team processes and outcomes would be necessary. Measuring and assessing collective processes and outcomes are difficult because they result from the interactions of individuals over time (Marks, Mathieu & Zaccaro, 2001). As such, it is necessary to control interactions during ITS training to assure desired levels of both team and individual performance (Bonner, Slavina, MacAllister, Holub, Gilbert, Sinatra, Dorneich & Winer, 2016).

The focus of assessment refers to either an outcome or a process being measured. Shared mental models may represent an example of an outcome insofar as shared mental models are knowledge structures that result from general and specific team interaction (Cannon-Bowers, Salas & Converse, 1993). That said, collective processes are important for collective outcomes. For example, shared mental models (outcome) cannot be constructed without sufficient levels of interaction (process). Some traditional ITS parameters may account for team processes, but more work is needed to map those processes to measurable actions or interactions (Burke, Feitosa & Salas, 2015). It is clear that both collective outcomes and collective processes need to be assessed in ITSs to sufficiently develop a learner model. The relative importance of each construct in any given ITS application will be determined not only by the training objectives of the ITS but also by the technical characteristics of the ITS that allow for the measurement of collective processes and outcomes.

Assessing collective processes (and collective outcomes) can be difficult in automated digital environments like ITSs (Seibert, et al. 2011). Most collective processes involve monitoring and interaction behaviors that are currently best captured by observer ratings (Marks et al., 2001). Assessing collective outcomes (e.g., responses) could be automated in ITSs, but effectively constructing collective-learner models would rely on an understanding of the processes that produced those outcomes as well. Using scripts to control and measure processes may facilitate automation in ITSs (Strijbos, 2011), but such an approach would limit the complexity of skills trained in ITSs. Thus, a significant challenge for collective ITSs is automating performance measurement in the digital training environment.

Discussion

The objective of the work reviewed here was to assess collective skills in the context of Army aviation within a simulated environment, such that the level of assessment, focus of assessment, and nature of the environment interacted to determine what should be assessed and how it could be assessed. Since the goal was to create assessments to provide feedback to improve collective performance, the worked focused not only on mission outcomes, but also on the processes that supported realization of those outcomes. Complete details can be found in published research reports, including measure definitions, validation results, and design of associated assessment tools (e.g., Seibert, et al., 2011; Bink, Dean, Ayers, Zeidman, 2014). Here, the focus is on findings concerning what could and could not be assessed regarding processes and outcomes, and given that, implications for the Generalized Intelligent Framework for Tutoring (GIFT) and related ITS environments.

More specifically, the work focused on improving learner feedback in the context of training exercises that were conducted in a networked virtual environment at the US Army Aviation Warfighting Simulation Cen-

ter (AWSC) at Fort Rucker, AL. The AWSC consists of networked cockpit simulators that can be reconfigured to represent various operational helicopters. The focus was on flight teams (teams of pilots coordinating across multiple helicopters) performing typical missions that attack weapons teams (AWTs) and scout weapons teams (SWTs) train and experience in combat. These missions required coordination within flight teams as well as with tactical operations centers (TOCs) and ground commanders, along with the identification, detection and engagement of targets. Hence, the tasks addressed were collective and focused on teamwork in addition to individual skills. Of interest were the processes (e.g., coordination and communication) that supported the achievement of outcomes (e.g., prosecution of targets).

Measures were developed using the Competency-Based Measures For Performance Assessment Systems (COMPASS) process (MacMillan, Entin, Morley & Bennett, 2013), which consisted of an iterative series of three workshops with subject-matter experts (SMEs) to develop and initially validate performance measures. First, a set of critical tasks was defined that were relevant for AWTs and SWTs. Next, descriptive indicators of high, average, and low performance on these tasks and underlying skills were created. Finally, measures were developed to quantify task performance to facilitate systematic learner feedback. Definitions included those for items to be captured manually via observers (observer-based) and automatically via simulator data streams (systems-based). SMEs who participated in the workshops were from diverse professional, civilian, and military backgrounds including military aviators, simulation training experts, and software engineers. SMEs represented the US Army Aviation Center of Excellence: the Directorate of Simulation (DOS), the Training and Doctrine Command Capability Manager (TCM) for Reconnaissance-Attack (RA), and the Aviation Captain's Career Course.

Overall, this process resulted in set of approximately 100 observable behaviors that captured collective performance during critical events. Based on these observable behaviors, a total of 115 observer-based measures that could discriminate high-performing from low-performing teams and that could be used to provide feedback were developed. In addition to the 115 observer-based measures developed in this effort, 33 additional system-based measures were defined using simulator data available during ATX. The critical question is what made these various measures likely, or not likely, to be successfully measured from simulator data feeds.

To answer this question, it is first necessary to consider what is in the simulator data stream that can be used to assess behavior. In this case, we investigated distributed interactive simulation (DIS) data log files produced during exercises at the AWSC. The files were based on standards from the Institute of Electrical and Electronics Engineers Standards for DIS Application Protocols (Institute of Electrical and Electronics Engineers, 1996) and the Simulation Interoperability Standards Organization Enumeration and Bit Encoded Values for use with Protocols for DIS Applications (Simulation Interoperability Standards Organization, 2006). The definitions for protocol data units (PDU) were also obtained, which are data messages that are exchanged on a network between simulation applications. Together, these items served as reference materials for the purpose of creating system-based measures that could be calculated automatically. In general, our analyses revealed that system-based data can be used to extract measures such as timing of events or success of an attack while observer-based data can better provide insights that are not easily obtained from system-based data alone. This distinction, therefore, poses limitations for automated tutoring systems focused on collective skills. The implication is that while some aspects of collective skills can be measured, including outcomes (e.g., success of an attack) and even aspects of process (e.g., sensor coordination when targeting – see the following example), some elements of process will be difficult to automatically measure (e.g., verbal, open-ended collaboration).

In particular, data were widely available for actions taken directly interacting with the simulator such as those involving aircraft movement, sensors, and targeting systems. One such measure, for instance, addresses confirmation of a target with appropriate marking technique given unit standard operating procedures. This measure reflects collective behavior because it is about coordination of the flight team with the

ground commander. In this case, the systems-based measure reflects key actions required for target confirmation that involve interaction with electronic systems, including indications of the position of the target and position of the laser designator (e.g., required data include "data from electromagnetic emission PDU; laser designator; position of target; position of laser designator"). This action can be assessed as correct or incorrect based on acceptable performance ranges, and hence, has the potential to be fully automated. However, it is important to note that this item alone does not reflect the full richness of the desired collective behavior. In this case, the complementary observer-based measure contrasted the simple action of "flight marks target" (acceptable performance) with the addition of a discussion about preferences concerning the marking strategy (superior performance defined as "flight discusses marking strategy with ground; marks target appropriately"). While the system data alone could provide evidence that a communication occurred, the quality of that exchange in context of a discussion regarding strategy was available only via observer. In this case, system data alone, and hence fully automated assessment, was insufficient.

Building further upon this type of distinction, of the 33 systems-based measures that were defined, most were viewed as being possible to implement with relative ease based on the then available technologies given the content of the data streams and functionality of the simulators. As another example of a possible item, the measure of distance to the wingman was computable in an automatic and continuous fashion, again reflecting coordination across the team within a well-defined standard for assessment. However, it should be noted that of the 33 system-based measures defined, 7 of them were ultimately identified as being unlikely to be fully automated in the foreseeable future. In general, while these items were typically about communications that could be automatically recorded and identified as having occurred, they were also associated with more open-ended collaboration where quality was somewhat ill defined or format was more open as multiple individuals interacted. For instance, one measure addressed coordination within the flight team in a short and concise manner that addressed a variety of variables (e.g., loitering, threats, approaches…). Similarly, another measure addressed the quality of collaboration during mission briefing in terms of covering all elements sufficiently in a timely manner. In fact, of the 7 measures, 5 focused on open-ended coordination and communication. The remaining measures concerned data relating to aspects of performance standards that were not fully defined at the time, and another measure that reflected focus of attention to a data feed (level of attention vs. just having the feed active), which would require additional technology to fully measure.

In summary, the primary limitation that resulted in this small set of fully implementable, automated items was the absence of key data in the available data stream that forced reliance on an observer. While data were widely available for actions taken directly interacting with the simulator such as those involving aircraft movement, sensors, and targeting systems, some elements of communication and coordination were unable to be fully captured. In many respects, these data reflect that some domain-specific tactics, techniques, and procedures may be more easily measured and assessed relative to more universal, qualitative collaboration oriented skills. Given this distinction, the strategy followed in the work reviewed was to divide assessment between those items that could be assessed automatically versus those that required observers. Although the assessment package was not "instructorless," the strategy did demand less from the instructor by automating where possible.

Recommendations and Future Research

What then, do these findings mean for the development of GIFT, or more generally, ITSs for collective skills? First, we believe these data suggest that a natural limit to what can be the focus of automated tutoring is the extent to which behavior can be assessed based on available data, and that this limit will be different for collective versus individual tasks (level of assessment). Second, even with data, within the context of collective skills, assessments will be hard to develop to fully support automated tutoring, in particular, those focused on aspects of process like open-ended collaboration (focus of assessment). Third, these limitations

suggest a strategy in that for collective skills, automated tutoring be focused on data-rich skills surrounding domain-specific tactics, techniques, and procedures, while focusing human instruction on more subtle aspects of open ended coordination and collaboration.

Based on the research presented here, and other similar efforts across a range of applications (different aircraft, games, etc.), our first recommendation is to always consider the assessment needs for the training *before* and during the design and implementation of the training system to identify the basic data points that will be required to inform and feed the assessment tools and/or ITSs. In general, there is a need to advance the overall capability of assessment tools for team-level skills such as communication and coordination. However, these advances are less about the technology and more about the content and volume of data that is required to assess those skills. Historically, data that are published within a networked training environment are specifically designed to represent entities and how those entities impact the environment in which they are situated (e.g., a weapon fire and damage assessment). Yet, data that would inform team dynamics are most often not published and therefore not available to assessment tools that would collect and interpret their impact. Limits on bandwidth and processor speed are no longer a major concern, and so the focus should be placed on identifying and publishing a richer data set on which collective assessment may be made. In general, better ITSs will be possible with more complete and targeted assessment opportunities through sophisticated analyses as technologies evolve (e.g., natural language processing).

Specifically, in the collective training arena that this research was undertaken, the combination of size of the training event (many interacting individuals in simulated aircraft, TOCs, etc.), the training objectives, and the data available within the environment placed limitations on what could be automatically assessed. For this reason, an ITS inside of the AWSC could only do so much good in the current and near-future configurations of the facilities. Our second recommendation therefore focuses on using automated measurement and assessment to alert and guide human observers and trainers during the training sessions to critical and specific times and places to view and observe detailed interactions between trainees and teams (e.g., arrival of aircraft at a critical location immediately prior to check in with the ground commander). Capitalizing on the items that can be computed automatically from the training environment, a system could cue observers and trainers to the most important interactions that they should be watching and assessing for the key critical interactions (e.g., the discussion that would then take place after arrival). Especially in events at the AWSC like those noted previously, where there are hundreds of trainees and only a handful of observers, this type of tool would facilitate more targeted and accurate assessment of collective performance and reduce the workload of the instructor/observers. These detailed assessments made by human observation could then be fed back into the training environment and used to trigger certain ITS behaviors, or be used to complement systems-based items on which the ITS acts. Combined, these elements could provide a rich set of ITS-fueled interventions.

Our final recommendation revolves around advancing the research around measuring collective performance, as well as the tools and strategies used to communicate that feedback to the team. As stated earlier, aggregated individual performance is not equivalent to collective performance. Therefore, we propose research into software models that may learn, measure, and then predict the impact of individual tasks on the overall performance of the training objective. The models may be able to measure and then assess the impact that an individual had on the collective mission. That information could be used to tailor and target individual intervention during the training by an ITS, in addition to informing the After Action Review. The concept of how an individual impacts the overall assessment of the team leads to additional research areas around the presentation and delivery of feedback to a team. More research is needed into what type of feedback should be given to teams in the context of ITSs, when should it be delivered, and how individuals within the team could then be provided with their own tailored evaluations.

Acknowledgments

We thank the US Army Aviation Center of Excellence, the Directorate of Simulation (DOS), the Training and Doctrine Command Capability Manager (TCM) for Reconnaissance-Attack (RA), and the Aviation Captain's Career Course for their earlier contributions to the efforts reviewed here. Likewise, we thank Melinda Seibert, John Stewart, Troy Zeidman, Courtney Dean, Aptima, Inc., and Imprimis, Inc.

References

Bink, M.L., Dean, C., Ayers, J. & Zeidman, T. (2014). *Validation and evaluation of Army aviation collective performance measures* (Research Report 1972). Ft. Belvoir, VA. US Army Research Institute for the Behavioral and Social Sciences.

Bonner, D., Slavina, A., MacAllister, A., Holub, J., Gilbert, S., Sinatra, A., Dorneich, M. & Winer, E. (2016). The hidden challenges of team tutor development. In R. Sottilare & S. Ososky (Eds.), *Proceedings of the fourth Generalized Intelligent Framework for Tutoring users symposium.* Orlando, FL: US Army Research Laboratory.

Burke, C. S., Feitosa, J. & Salas, E. (2015). The unpacking of team models in GIFT. In R. Sottilare & A. Sinatra (Eds.), *Proceedings of the third Generalized Intelligent Framework for Tutoring users symposium.* Orlando, FL: US Army Research Laboratory.

Cannon-Bowers, J. A., Salas, E. & Converse, S. A. (1993). Shared mental models in expert team decision making. In N. J. Castellan, Jr. (Ed.), *Current issues in individual and group decision making* (pp. 221–246). Hillsdale, NJ: Erlbaum.

Department of the Army (2012). *Training development in support of the operational domain*. TRADOC Pam 350-70-1. Fort Eustis: Headquarters, United States Army Training and Doctrine Command.

Fletcher, J. D. & Sottilare, R. A. (2013). Shared mental models of cognition for intelligent tutoring of teams. In R. Sottilare, A. Graesser, X. Hu & H. Holden (Eds.), *Design recommendations for intelligent tutoring systems. Volume 1: Learner models.* Orlando, FL: US Army Research Laboratory.

Fletcher, J.D. & Sottilare, R.A. (2017). Shared Mental Models in Support of Adaptive Instruction for Team Tasks using the GIFT Tutoring Architecture. *International Journal of Artificial Intelligence in Education.* DOI: 10.1007/s40593-017-0147-y.

Gilbert, S., Winer, E., Holub, J., Richardson, T., Dorneich, M. & Hoffman, M. (2015). Characteristics of a multi-user tutoring architecture. In R. Sottilare & A. Sinatra (Eds.), *Proceedings of the third Generalized Intelligent Framework for Tutoring users symposium.* Orlando, FL: US Army Research Laboratory.

Hmelo-Silver, C. E., Chernobilsky, E. & Jordan, R. (2008). Understanding collaborative learning processes in new learning environments. *Instructional Science, 36*, 409–430.

Institute of Electrical and Electronics Engineers (1996). *Standards for distributed interactive simulation application protocols* (Std 1278.1-1995). New York, NY: Institute of Electrical and Electronics Engineers, Inc.

MacMillan, J., Entin, E. B., Morley, R. M. & Bennett Jr., W. R. J. (2013). Measuring team performance and complex and dynamic military environments: The SPOTLITE method. *Military Psychology,25,* 266–279.

Marks, M. A., Mathieu, M. & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review, 26*, 356–376.

Salas, E., Sims, D. E. & Burke, C. S. (2005). Is there a "Big 5" in teamwork? *Small Group Research, 36*, 555–599.

Seibert, M. K., Diedrich, F. J., Stewart, J. E., Bink, M. L. & Zeidman, T. (2011). *Developing performance measures for Army aviation collective training* (Research Report 1943). Arlington, VA. US Army Research Institute for the Behavioral and Social Sciences.

Simulation Interoperability Standards Organization (2006). *Simulation interoperability standards organization enumeration and bit encoded values for use with protocols for distributed interactive simulation applications* (Ref 010-2006). Orlando, FL: Simulation Interoperability Standards Organization, Inc.

Sottilare, R. A. (2013). Pushing and pulling toward future ITS learner modeling concepts. In R. Sottilare, A. Graesser, X. Hu & H. Holden (Eds.), *Design recommendations for intelligent tutoring systems. Volume 1: Learner models.* Orlando, FL: US Army Research Laboratory.

Sottilare, R., Holden, H., Brawner, K. & Goldberg, B. (2011). Challenges and Emerging Concepts in the Development of Adaptive, Computer-based Tutoring Systems for Team Training. In Proceedings of the Interservice/Industry Training Systems & Education Conference, Orlando, Florida, December 2011.

Sottilare, R.A., Brawner, K.W., Goldberg, B.S. & Holden, H.K. (2012). The Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: US Army Research Laboratory, Human Research & Engineering Directorate (ARL-HRED).

Strijbos, J. (2011). Assessment of (computer-supported) collaborative learning. IEEE Transactions on Learning Technologies, 4, 59–73.

CHAPTER 26 – Cognitive Assessment as Service in the Generalized Intelligent Framework for Tutoring (GIFT)

Xiangen Hu[1,3], Sheng Xu[3], Robert Sottilare[2], and Dietrich Albert[4,5]
University of Memphis[1], US Army Research Laboratory[2], Central China Normal University[3],
Graz University of Technology[4], University of Graz[5]

Introduction

One of the most common behaviors observed in various forms of intelligent tutoring system (ITS) implementations is similar to the typical stimulus-response (S-O-R) paradigm (https://www.quora.com/What-is-the-S-O-R-model-about). In any sequence of ITS tutoring sessions, the ITS presents a stimulus (such as seed question or a scenario) to the learner. The learner's *response* will then be a *stimulus* to the ITS. To further illustrate such a turn-alternating S-R behavior observed in an ITS, consider a typical ITS-learner interaction sequence:

A. …,...

B. The ITS presents a scenario to the learner.

C. The learner tries to understand the scenarios (as stimulus) and responds to the scenarios (as response).

D. The ITS evaluates the learner response (as stimulus) and presents a follow-up scenario to the learner (as a response).

E. …,...

In this outlined sequence of micro-actions between the ITS and learner, the most important action for the learner is to *understand* the scenarios presented by the ITS. The most important action for the ITS is to *evaluate* the responses by the learner and respond to the learner's contribution.

In current and earlier volumes of the series, more detailed analyses of ITSs have been presented. In general, an ITS has been considered as an integrated collection of more complicated models and processes (Sottilare, Graesser, Hu & Holden, 2013; Sottilare, Graesser, Hu & Goldberg, 2014; Sottilare, Graesser, Hu & Brawner, 2015; Sottilare, Graesser, Hu, Olney, Nye & Sinatra, 2016). With the limited space and restricted focus of the current chapter, we focus only on the evaluative component of ITSs (step D of the outline sequence), especially, we consider such component in the context of the General Intelligent Framework for Tutoring (GIFT). For the purpose of illustration, we consider two types of learner's responses: categorical responses (such as multiple choices) and natural language responses. For each of these two types of response types, we present theory-based assessment models, implemented as standalone software service following the best practice of service-oriented architecture (SOA). We argue that such web services can be used to serve GIFT and as an example of cognitive assessment web service for GIFT.

Related Research

We first introduce two assessment frameworks for the two types of learner's responses. The first is called Multinomial Processing Tree (MPT) models and second is called Semantic Representation and Analysis

(SRA). This chapter only addresses the two example frameworks at the conceptual level and does not go into the mathematical details.

Multinomial Processing Tree (MPT) Models

First, we introduce MPT models by considering an example of the typical Urn model (https://en.wikipedia.org/wiki/Urn_problem) with two types of balls: r red (R) balls and b blue (B) balls (Figure 1). If one randomly draw two balls from the urn (in sequence, one ball at a time), What is the probability of that the first ball is blue (B) and second ball is red (R)? What would be the model to describe the expected probability for the outcomes?

The probability of observing event (B, R) would be $p(1-q)$ where $p = \frac{b}{b+r}$ and the value q would be different depending on the drawing method: with or without replacement. If the drawing is without replacement, then $q = \frac{b}{b+r-1}$; if the drawing is with replacement, then $q = \frac{b}{b+r}$.

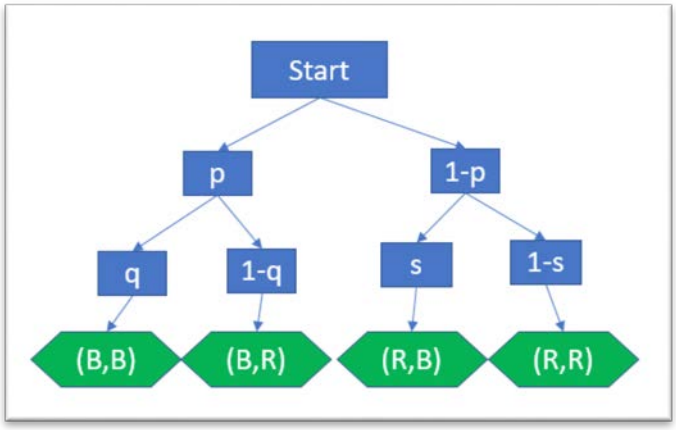


Figure 1. MPT for the Urn model.

An MPT model for the urn example would be

- $p\{(B,B\} = pq$,
- $p\{(B,R)\} = p(1-q)$,
- $p\{(R,B)\} = (1-p)s$, and
- $p\{(R,R)\} = (1-p)(1-s)$

where q and s are different. If the drawing is with replacement, then ($q = s = \frac{b}{b+r}$); if the drawing is without replacement, then ($q = \frac{b}{b+r-1}$, $s = \frac{b-1}{b+r-1}$).

The model described here is a typical MPT model. It is a multinomial model and there are special parameters (such as p, q, and s) in the model that can be interpreted in the context of a "tree-like" structure. Multinomial models with these two properties are called MPT models. It has been shown that any multinomial model can be reparametrized into an MPT model (Hu & Batchelder, 1994).

The intuitive and simplicity of MPT models makes it possible to apply in analyzing experiments in cognitive psychology and other areas of research (Batchelder & Riefer, 1999). In the past 30 years, MPT modeling has been used as one of the formal approaches to measuring cognitive processes, such as the capacity to store and retrieve items in memory, make inferences and logical deductions, or discriminate and categorize similar stimuli. While such processes are not directly observable, theoretically they can be assumed to interact in certain ways to determine observable behaviors. The goal of multinomial modeling is to identify which underlying factors are important in a cognitive task, explain how those processes combine to create observable behavior, and then use experimental data to estimate the relative contributions of the different cognitive factors. In this way, multinomial models can be used as tools to measure unobservable cognitive processes (Batchelder, Hu & Riefer, 2013).

In the next section, we illustrate the utility of MPT models as an assessment tool by giving a relevant example in cognitive psychology. Especially, MPT models are used to analyze categorical responses to infer underlying unobservable cognitive capacities. We then describe how MPT modeling can be used as an assessment service for GIFT.

Example 1: Application of MPT Modeling in Assessing Cognitive Ability

From the perspective of mathematical statistics, multinomial models are developed for categorical data, where each participant's response falls into one and only one of a finite set of observable data categories. When these data come from a cognitive experiment, each participant in an experimental group produces a categorical response to each of a series of items, for example, pictures are "recognized" or "not recognized" or letter strings are judged to be "words" or "non-words". In the context of learning/assessment environments, learners are frequently asked to make multi-alternative forced choices, such as multiple choice (MC). Even in non-formal assessment/testing situations, learner's learning behaviors can be classified into mutually distinctive behavior categories. All these data are suited for MPT models.

Most data sets for multinomial modeling involve more than two response categories. There also may be more than one type of item, each with its own system of response categories. For example, in a source-monitoring experiment, participants study a list of items from two sources, Source 1 or Source 2 (e.g., presented by a reliable vs. an unreliable learning source). Later, participants are given a recognition memory test consisting of three types of items, namely, the two types of old list items and new distracter items, and they must classify each tested item as Source 1, Source 2, or New. The resulting multinomial data structure consists of three category systems, each with three response categories. If the responses in different category systems are independent and category counts within a system follow a multinomial distribution, the probability of the data structure is given by the product of three multinomial distributions, one for each category system. The MPT model for the source-monitoring experiments is shown in Figure 2.

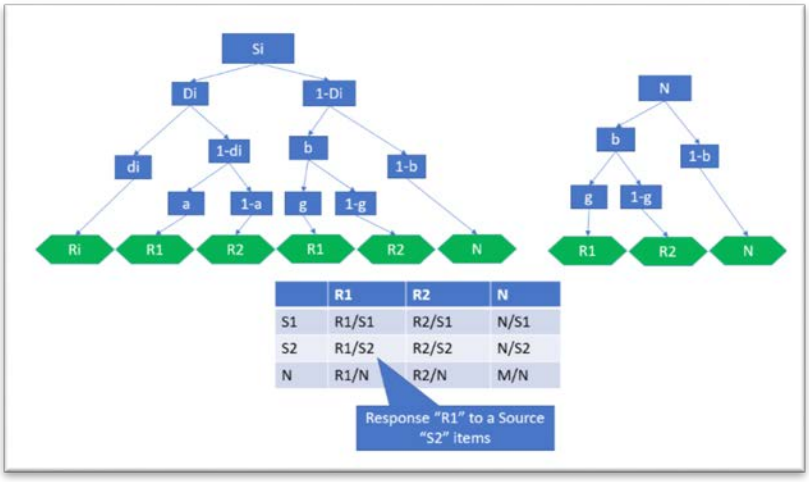


Figure 2. Model and data structure for the source-monitoring experiment (Batchelder & Riefer, 1990), where i takes the value of 1 and 2 in the symbol Si, Ri, Di, and di.

In this model, there are total of 7 parameters, *D1*, *D2*, *d1*, *d2*, *b*, *a*, and *g*. All the parameters are probabilities and used as measurement of underlying cognitive abilities:

- *D1* and *D2* are for the detection of a test item as old item.
- *d1* and *d2* are for the discrimination of test items given they are detected as old items.
- *b* is the response bias for an item when failed to detect as old,
- *a* is response bias to source 1 item when the item is detected as an old item.
- *g* is response bias to source 1 item when the item is not detected as an old item.

The mathematical form of the MPT model for source-monitoring is specified in terms of categorical probabilities:

- $p\{R1/S1\} = D1d1 + D1(1-d1)a + (1-D1)bg,$
- $p\{R2/S1\} = D1(1-d1)(1-a) + (1-D1)b(1-g),$
- $p\{N/S1\} = (1-D1)(1-b)$
- $p\{R1/S2\} = D2(1-d2)(1-a) + (1-D2)b(1-g),$
- $p\{R2/S2\} = D2d2 + D2(1-d2)(1-a) + (1-D2)b(1-g),$
- $p\{N/S2\} = (1-D2)(1-b)$
- $p\{R1/N\} = bg,$
- $p\{R2/N\} = b(1-g),$

- $p\{R3/N\} = (1 - b)$

The construction of the model is intuitive both from the point of the tree structure and from analysis of the behavior of the response. For example, $p\{R1/S2\}$ is the probability of given a source 2 item the participant responds incorrectly as source 1 item. The probability is the sum of two terms:

- $D2(1 - d2)(1 - a)$: Detected as old item ($D2$), but could not discriminate $(1 - d2)$, but guess the item as Source S2 $(1 - a)$.

- $(1 - D2)b(1 - g)$: Fail to detect it $(1 - D2)$, biased it to old (b), and guess it as source S2 $(1 - g)$.

With this model, we can measure unobservable underlying capacities such as memories for content (*D1*, *D2*) and for context (source of information *d1, d2*). In addition, different types of responses (*b*: response bias at the level of content information; *a, g*: response bias at the level of context information) are also measured.

In general, assuming in an event there are multiple observed behavior categories. An MPT model links the observed categorical probability $\Pr(C_j)$ and underlying capacities (in the form of probabilities) θ_{ks} in the form of a general (non-linear) mathematical form ((Hu & Batchelder, 1994):

$$\Pr(C_j) = \sum_{i=1}^{I_j} c_{ij} \prod_{k=1}^{K} \left[\prod_{s=1}^{S_k} \theta_{ks}^{\alpha_{ijks}} \right]$$

where $\sum_{j=1}^{J} Pr(C_j) = 1, \sum_{s=1}^{S} \theta_{ks} = 1, c_{ij} \geq 0, \alpha_{ijks} \geq 0, s = 1, \dots, S, k = 1, \dots, K, i = 1, \dots, I_j, j = 1 \dots, J$. This general mathematical form provides a general assessment framework for assessing underlying cognitive capacities from observed categorical behavior responses. Next we provide an example where GPT model is used to analyze categorical responses from a typical multiple-choice question

Example 2: Application of MPT Modeling in Analyzing Multiple-Choice Questions

MPT models have been used to analyze responses in multiple-choice questions (Batchelder & Riefer, 1999). To illustrate how MPT model would be used, we consider the following example.

Two metal balls are the same size but one weighs twice as much as the other. The balls are dropped from the roof of a single-story building at the same instant of time. The time it takes the balls to reach the ground below will be:

A. About half as long for the heavier ball as for the lighter one.

B. About half as long for the lighter ball as for the heavier one.

C. About the same for both balls.

D. Considerably less for the heavier ball, but not necessarily half as long.

E. Considerably less for the lighter ball, but not necessarily half as long.

When air resistance is not important, C is the right answer. But all other choices are wrong due to the following misconception: *When air resistance is not important, objects of different masses fall at different*

rates. An MPT model that models the relationship between response categories and the underlying misconceptions can be constructed as Figure 3.

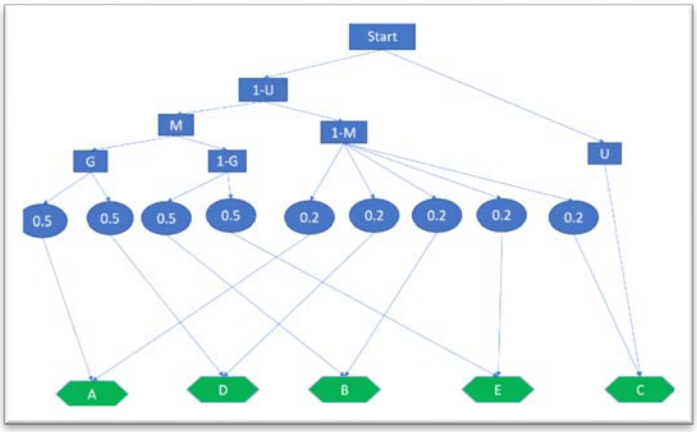


Figure. 3. MPT models for the selected sample Force Concept Inventory (FCI) question.

In Figure 3, U is the probability of the students understand that "*When air resistance is not important, objects of different masses fall at same rates*", M is the probability that students have the misconception "*When air resistance is not important, objects of different masses fall at different rates*", and G is the probability that students know some property about gravity.

With these notations and assume equal guessing rate, we have the following categorical probabilities:

- $P(A) = 0.5(1-U)MG + 0.2(1-U)(1-M)$
- $P(B) = 0.5(1-U)M(1-G) + 0.2(1-U)(1-M)$
- $P(C) = U + 0.2(1-U)(1-M)$
- $P(D) = 0.5(1-U)MG + 0.2(1-U)(1-M)$
- $P(E) = 0.5(1-U)M(1-G) + 0.2(1-U)(1-M)$

This model has five observable categories (4 degrees of freedoms in the data) and three parameters. Goodness-of-fit can be tested with Chi-square (df=1).

MPT Models as General Assessment Modeling

Since the 1990s, MPT modeling has become an increasingly popular approach to cognitive modeling. Its use has been facilitated by several software packages that can perform parameter estimation and hypotheses testing. To date, there have been over a hundred examples of the application of MPT modeling. Most of these applications have been in the standard cognitive areas of memory, reasoning, and perception; additional applications can be found in clinical, social, and developmental psychology (Batchelder & Riefer, 1999; Erdfelder, 2009). There are also research that explore the statistical properties of these models. For example, Hu and Batchelder (1994) formulated the MPT models into a general mathematical model. Klauer and his group created hierarchical MPT models to handle variation in parameter values due to individual

differences in the participants (Klauer, 2010). There are latent class MPT models that can be used to model subgroups of participants with different cognitive abilities (Matzke, et al 2015).

Parallel to mathematical, statistical, and empirical studies of MPT models, one of the enabling factors that makes the MPT model an increasing analytical and assessment framework is the availability of software implementations. Since the very earliest versions of source.exe, mbt.exe, and gpt.exe (Hu & Phillips, 1999) there have been several new implementations of MPT modeling software. Moshagen (2010, Table 1) has listed and reviews a few landmark implementations of software packages.

There has been increasing popularity of R in the research community. Consequently, there are two R-packages designed for MPT models:

- *MPTinR* (https://cran.r-project.org/web/packages/MPTinR/index.html) and
- *mpt* (https://cran.r-project.org/web/packages/mpt/index.html).

We have developed a simple web service that interfaces with MPTinR (see Appendix A). Such web service accepts input model information in the form of JavaScript Object Notation (JSON). Next, we explore the possibility of using MPT modeling as an assessment framework for the Engine for Management of Adaptive Pedagogy (eMAP)

MPT Modeling as Assessment Service for GIFT eMAP

The Adaptive CourseFlow in GIFT uses a mechanism called eMAP to dynamically present learning resources to the learner based on learner state attributes of cognitive knowledge, affective state, and cognitive skill. One of the most important phases in setting up Adaptive CourseFlow is Check on Learning. Check on Learning is most often achieved by assigning learners with multiple-choice questions. Assume MPT models can be created for the multiple-choice questions, similar to the example of Figure 3. An MPT web service (described in the Appendix A) would be used to analyze student's answers.

Semantic Representation and Analysis (SRA) Framework

We next provide another example of assessment framework based early work of the authors on the SRA framework, which is an established service in GIFT (see Hu, Nye, Gao, Huang, Xie, Shubeck, 2014).

Basic Assumptions of SRA

SRA provides a general framework for conceptualizing and applying existing semantic extraction/encoding methods, such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), Hyperspace Analogue to Language (HAL; Burgess, 1998), and bound encoding of the aggregate language environment (BEAGLE; Jones & Mewhort, 2007). The two key elements of SRA are the vector representation of the semantics of language entities (words, idioms, phrases, sentences, paragraphs, documents, etc.) and the numerical relations between language entities (such as similarity, relatedness, or semantic overlap). The basic requirements for SRA are that the representations must be language agnostic and computationally feasible. Hu, Cai, Graesser, and Ventura (2005) outlined SRA based on the following assumptions:

1) **Hierarchical Representation**: Various levels of a language entity may have their semantics represented differently. The basic language entities are hierarchical such as words, phrases, sentences, paragraphs, and documents.

2) **Algebraic Representation**: The semantics of any level of language entities must be capable of being represented numerically or algebraically. A number of semantic extraction/encoding methods (e.g., LSA) have been used to numerically represent semantics for all levels of language entities through the creation of semantic spaces. Most of these examples have the same mathematic representation for different levels of language entities, such as numerical vectors of a given dimension for word, sentence, or paragraph. This assumption allows them to be represented differently.

3) **Computational Aggregation**: The semantics of a higher-level language entity are computed as a function of semantics for its lower level language entities. Also, at the lowest level of language entities, a numerical semantic comparison measure must exist between any two items (e.g., words). For example, if words are in the form of vectors, **a**, **b**, **c**…, a sentence can be in the form of a function ■ of the word vectors, such as (**a**■**b**■**c**,…), where ■ can be vector summation (as it is in the case of LSA, for example).

These three assumptions are the foundation of a general framework underlying most existing semantic extraction/encoding methods. The hierarchical assumption and the algebraic representative assumption work together to ensure that the language entities can be computed mathematically. This final assumption emphasizes the idea that comparisons occurring at the most basic level can be inputs for higher levels (e.g., the similarity of paragraphs can be computed from the similarity between their constituent sentences). It is important to note that the popular encoding/decoding methods such as LSA are special example of the SRA framework. Within the SRA framework, induced semantic structure (ISS) is obtained.

Induced Semantic Structure (ISS)

An important concept derived from the SRA framework is the ISS. ISS focuses on numeric relations between language entities while deemphasizing the encoding details (such as the vector representation) for the semantic spaces. ISS considers a target word and an ordered list of its top nearest neighbors in a semantic space (Hu et al., 2005). Introducing the concept of ISS in addition to the original semantic spaces made it possible for two features of SRA: first, any two semantic spaces with overlapping lexicons can be compared (Hu et al., 2005); and second, any set of texts (single term or a collection of terms) can be projected onto a customize set of domains. Relevant to the focus of the current chapter, we briefly describe the second feature. We call it semantic spectrum analysis (SSA).

Semantic Spectrum Analysis (SSA)

Spectrum analysis is a term borrowed from physics. Such analysis helps researchers to understand the basic elements in a physical entity such as piece of sound or a beam of light. SSA helps researchers to understand basic semantic elements contained in each piece of text. The necessary condition for SSA is to have an existing semantic representation and a set of pre-specified domains. For example, with a set of domains extracted from glossary.com (http://www.glossary.com), SSA can project the term "book" to each of the 20 domains:

environment: 2.33
home and garden: 4.51
local business: 3.41
computers and internet: 3.37
science and mathematics: 4.43
beauty and style: 2.44
news and events: 2.00
games and recreation: 4.12
health: 4.60
social science: 4.34
sports: 2.14
society and culture: 4.21
food and drink: 4.29
consumer electronics: 2.90
arts and humanities: 3.76
pets: 2.69
politics: 4.09
family and relationships: 4.50
cars and transportation: 1.56
travel: 2.39

The absolute numerical values associated with each of the categories are not directly interpretable without considering the specific semantic spaces used. In this example, the numerical values are the semantic associations of the term "book" to each of the domains using the semantic space created by the Touchstone Applied Science Associates (TASA) corpus. For any given piece of text, when semantic associations of each term to each of the "domains" are computed, the semantic association of the entire piece of text with the domains can be computed by simple aggregation. Figure 4 is a graphic display of so called *semantic decomposition*. Semantic decomposition is one of the assessment service provided by The *Domain Specific Semantic Processing Portal (DSSPP)*. DSSPP is the second example we use to demonstrate the cognitive assessment service for GIFT.

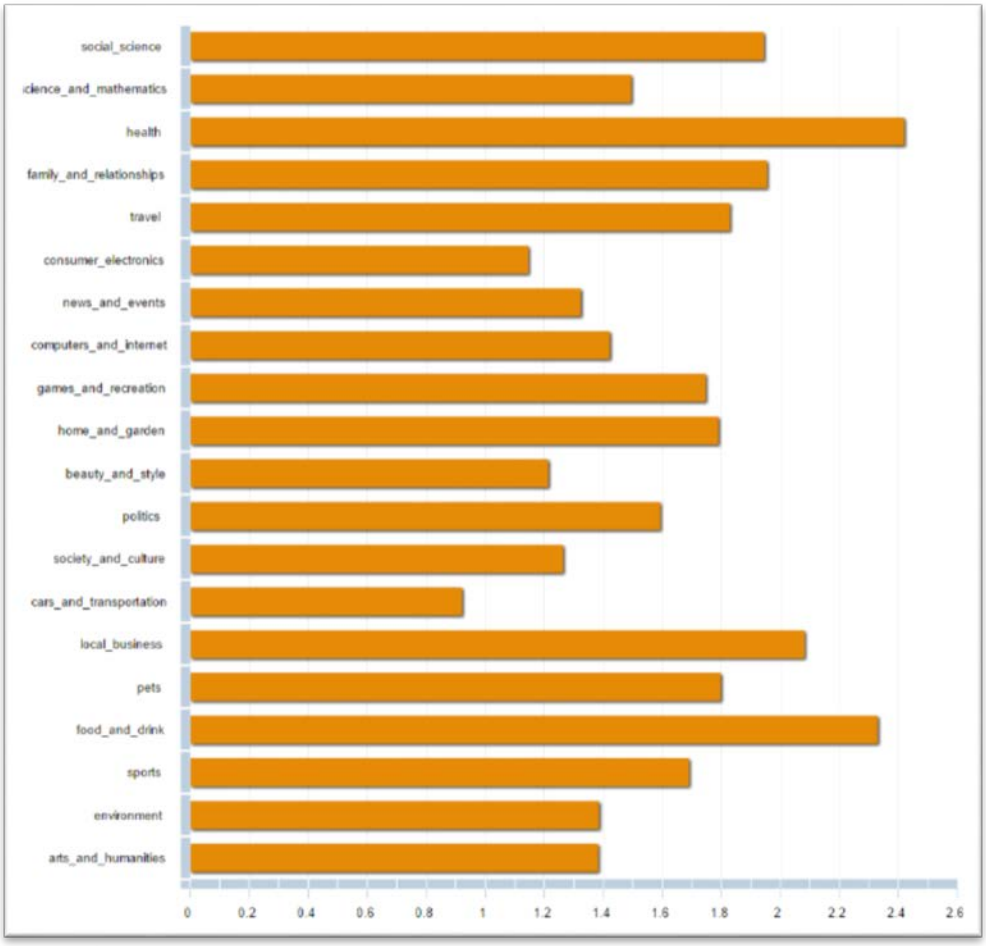


Figure. 4. Graphic display of semantic decomposition of a paragraph. The domains ("social sciences", "health", etc.) are extracted from http://www.glossary.com.

The Domain Specific Semantic Processing Portal (DSSPP) Assessment Service for GIFT

DSSPP (http://dsspp.skoonline.org) is a proof of concept implementation of SRA. DSSPP is implemented as a cloud based web service currently in the form of Google App Engine and Amazon Elastic Compute Cloud. DSSPP provides four web services (for details, see Appendix B):

1. ISS in the form of nearest neighbors for available semantic spaces. For any term (e.g., a word), DSSPP web service provides a list of associated terms with association strength and term weight (See Appendix B)

2. Semantic relations (e.g., similarity) between any two pieces of texts within or between two semantic spaces

3. SSA for any piece of text

4. Learner's characteristics curves (LCCs) (details see Hu & Martindale, 2008) for sequence of student's response to any given question and answer key. These functionalities can be used for several applications, including versions of AutoTutor (Nye, Hu, Graesser & Cai, 2014).

Of the four of the web services provided by DSSPP, 2–4 can be used as cognitive assessment service for GIFT.

Discussion

As we have pointed out at the start of the chapter, one of the most critical steps in ITSs is to evaluate students' contributions. Students contributions can be in the form of categorical responses (such as standard survey/test, well-defined category of behaviors) or non-categorical (such as verbal/natural language responses or ratings). This chapter presents two assessment frameworks that are implemented in the form of web services. The MPT modeling assessment service can be used to evaluate any categorical behavior observed in a GIFT-enabled ITS learning environment. For example, for some well-constructed multiple-choice GIFT survey/test question, a MPT model can be created (like the MPT model for the earlier physics example). After a group of responses from students, GIFT can send the model and data to an MPT assessment portal (as it is demonstrated in Appendix A) to obtain measures of students' capacity.

The web service of DSSPP can be used to evaluate students' verbal (NL) responses. For example, students NL input can be evaluated by comparing semantically with stored answers or comparing with previous inputs. Such evaluation can be achieved by a simple request DSSPP service to produce a single numerical value as semantic similarity (such as /comparetext, see Appendix B) or to produce a vector of numerical values (such as /ssa, see Appendix B). Such evaluations can be used as standalone web service to process single evaluation or integrated as part of sophisticated tutoring interactions such as AutoTutor's. The LCCs web service (/lcc, see Appendix B) can be used to created local student model to manage turn-by-turn interactions in GIFT-enabled ITS learning environment (Morrison, Nye & Hu, 2014).

Recommendations and Future Research

The assessment frameworks introduced in this chapter are only two of many potential assessment frameworks. There are other well-known examples such as Bayesian knowledge tracing models (Corbett & Anderson, 1995) can be used in the similar way. Most importantly, we propose to have cognitive assessment as service to GIFT as a functional module similar to other critical modules of GIFT, such as a pedagogical module or a domain module. An assessment module can be used to serve all assessment needs for GIFT. Having assessment service as a functional module makes it possible for GIFT to incorporate new research, development, and implementation of assessment methodology.

Current implementation of MPT assessment web service can only analyze aggregated frequency tables. There are extensions of MPT models that can analyze data from each response (Matzke, et al 2015). Successful implementation of this method will make it possible to assess students' contribution at the individual level.

GIFT provides highly adaptive and individualized tutoring environments. Current DSSPP considers only domain-specific semantic processing. A true assessment service for GIFT will need also be individualized. The next step to implement SRA is to add individualization and context sensitive assessments

References

Andrews, M. Vigliocco G. and Vinson D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychol Rev.* 2009 Jul: 116(3):463–98. doi: 10.1037/a00116261.

Batchelder, W. H., Hu, X. & Riefer, D. M. (2013). Multinomial modeling. In H. Pashler (Ed.). *Encyclopedia of Mind*. (pp. 538–541). Sage Publications Inc.

Burgess, Kurt. 1998. From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. In *Behavior Research Methods* 30(2):188-198. doi: 10.3758/BF03200643.

Corbett, A. T.; Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction.* 4 (4): pp. 253–278.

Erdfelder, E., Hilbig, B., Auer, T. & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Journal of Psychology*, 217, pp. 108–124.

Hu, X. & Batchelder, H. W. (1994). Statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*. 59 (1), pp. 21–47.

Hu, X., Cai, Z. Graesser, A. C., Ventura, M. (2005). Similarity between semantic spaces. Bara, B. G., Barsalou, L. & Bucciarelli, M. (Eds.). *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Stresa, Italy: Cognitive Science Society.

Hu, X. & Phillips, G. A. (1999). GPT.EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods, Instrumentation, and Computers*. 31 (2), pp. 220-234.

Hu, X. & Martindale, T. (2008). Enhancing learning with ITS-style interactions between learner and content. *Interservice/Industry Training, Simulation & Education*, 8218, pp. 1–11.

Hu X., Nye B.D., Gao C., Huang X., Xie J., Shubeck K. (2014). Semantic representation analysis: a general framework for individualized, domain-specific and context-sensitive semantic processing. In: Schmorrow D.D., Fidopiastis C.M. (eds) Foundations of augmented cognition. Advancing human performance and decision-making through adaptive systems. AC 2014. Lecture Notes in *Computer Science*, 8534. Springer, Cham.

Itagaki, M. Aue, A. and Aikawa, T. (2006). Detecting inter-domain semantic shift using syntactic similarity. In: Calzolari, N., Gangemi, A., Macgaard, B., Mariani, J., Odijk, J., Tapias, D. (eds.) 5th International Conference on Language Resources and Evaluation (LREC), pp. 2399–2402. European Language Resources Association, Paris.

Jones, M.N., and D.J.K. Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Pschol. Rev*. 104:1–37. 4.

Klauer KC. Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*. 2010; 75: pp. 7–98.

Landauer, T. & Dumais, S.T. (1997). A solution to Plato's Platform: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.

Matzke, D., Dolan, C. V., Batchelder, W. H. & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, 80, pp. 205–235.

Morrison, D. M., Nye, B. & Hu, X. (2014). Where in the data stream are we?: Analyzing the flow of text in dialogue-based systems for learning. In R. A. Sottilare, X. Hu, H. Holden & K. Brawner (Eds.) *Design Recommendations for Intelligent Tutoring Systems: Volume 2: Adaptive Instructional Strategies and Tactics* (pp. 217–223). US Army Research Laboratory.

Moshagen, M. (2010), multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, 42(1) pp. 42–54.

Nelson, D. McEvoy, C.L., Schreiber, T. (2004). The University of South Florida, word association, rhyme and word fragment norms. *Behavior Research Methods, Instruments & Computers*, 36, 408–420.

Nye, B., Hu, X., Graesser, A.C., Cai, Z. (2014). AutoTutor in the cloud: a service-oriented paradigm for an interoperable natural-language ITS. *Journal of Advanced Distributed Learning Technology*. 2(6), pp. 35–48.

Sottilare, R., Graesser, A., Hu, X. & Holden, H. (Eds.). (2013). *Design Recommendations for Intelligent Tutoring Systems - Volume 1 - Learner Modeling (1st edition)*. Orlando, Florida: US Army Research Laboratory. ISBN 978-0-9893923-0-3.

Sottilare, R., Graesser, A., Hu, X., and Goldberg, B. (Eds.). (2014). *Design Recommendations for Intelligent Tutoring Systems: Volume 2 - Instructional Management*. Orlando, FL: US Army Research Laboratory. ISBN 978-0-9893923-3-4.

Sottilare, R., Graesser, A., Hu, X., and Brawner, K. (Eds.). (2015). *Design Recommendations for Intelligent Tutoring Systems: Volume 3 - Authoring Tools and Expert Modeling Techniques*, Orlando, FL: US Army Research Laboratory. ISBN 978-0-9893923-7-2.

Sottilare, R., Graesser, A., Hu, X., Olney, A., Nye, B. and Sinatra, A. (Eds.). (2016). *Design Recommendations for Intelligent Tutoring Systems: Volume 4 - Domain Modeling*. Orlando, FL: US Army Research Laboratory. ISBN 978-0-9893923-9-6.

Appendix A: Implemented MPT Assessment Service based on MPTinR

R is installed on a cloud server (currently located on http://www.auto-tutor.com/) and simple interface is built to use MPTinR (https://cran.r-project.org/web/packages/MPTinR/index.html). Such a web service accepts MPT models in the form of JSON and computed output is also in the form of JSON. As an example, the specified model input and output parameter estimation can be seen from the shortened URL http://tinyurl.com/gq7kyc5, where both input and output are in the form of JSON and can be easily incorporated in GIFT programming environment.

Appendix B: Implemented DSSPP Assessment Service Based on SRA

DSSPP is currently implemented on a Google Cloud Server (http://dsspp.skoonline.org). The assessment portal provides the following services:

1) Induced semantic structure in the form of nearest neighbors with term weights and association strength. For example, the end point of /base will return the nearest neighbor for any given term, if appropriate parameters are provided (permanent short link is here: https://goo.gl/XQ0hbo).

2) Semantic relations (e.g., semantic similarity) between any two pieces of texts in a given semantic space is a number between 0 and 1, where 0 means no semantic relation and 1 means perfect semantic similarity. An example can be seen by using this link: https://goo.gl/DRFG2b.

3) Semantic Spectrum Analysis (SSA) for any piece of text are in the form of a numerical vector where each elements of the vectors is the semantic association of the target text and a pre-defined domain, such as environment or health (example of SSA can be seen from this permanent short link: https://goo.gl/V3wQjG).

4) Learner's Characteristics Curves (LCC) for sequence of student's response to any given question and answer key (Hu & Martindale, 2008). For given target text (as answer key), each student's contribution (answer) is decomposed into six different components: current contribution (CC), total contribution (TC), relevant new (RN), irrelevant new (IR), relevant old (RO), and irrelevant old (IO). This can be tested by using a short link https://goo.gl/yoTqJ0.

5) Definitions of the parameters in DSSPP web services.

    - text - the text to parse (for /base).

- domains - the list is a space-separated domain (no domain has a space in it).
- SS - the semantic space to use.
- type - an integer for the grouping method to use 1-assn, 2-rankby, 3-weight.
- guid, userGuid - a unique identifier for the account of the user (do not change).
- text1, text2 are the texts to compare (for /comparetext).
- minStrength is the lowest cosine to use (default 0.0).
- minWeight is the lowest weight to use (default 0.0).
- format is either xml or json (default “xml”).
- id1, id2, and notes are used (for /comparetext).
- sort_method is an integer 0-rankby, 1-cosine, 2- weight (default 0).
- wc: weight criteria for input text (default=0).
- etop: number of top nearest neighbors for each term (default=10).
- ttop: number of top nearest neighbors for total terms combined (default=50).
- notes: Any text user may use as notes.
- target: target for to compare (as answer key, for /lcc).
- current: current input of student (as student answer, for /lcc).

CHAPTER 27 – Assessment in AutoTutor

Zhiqiang Cai[1], Arthur C. Graesser[1], Xiangen Hu[1,3], and Bor-chen Kuo[2]
University of Memphis[1], National Taichung University of Education[2], Central China Normal University[3]

Introduction

AutoTutor is a type of intelligent tutoring system (ITS) that uses natural language conversation to help students learn (Graesser, 2016; Graesser, Forsyth & Foltz, 2016; Graesser, Person, Harter & Tutoring Research Group, 2001; Zapata-Rivera, Jackson & Katz, 2015). In a traditional classroom learning environment, a teacher delivers knowledge to a group of students in a linear way. That is, all students in the class follow the same learning path as the teacher plans. Computer tutoring makes it possible to allow each student to go through one's own learning path that maximizes learning and minimizes the cost. ITSs are designed to meet these objectives (Sottilare, Graesser, Hu & Goldberg, 2014).

ITSs are capable of adaptively selecting learning objects to deliver to an individual learner. The selection mechanism is basically a mapping from learning objects to learners' profiles. In this chapter, learning objects refer to digital entities that may be used for learning, education, or training. They could be of very different grain sizes. A learning object could be as large as an entire knowledge domain, such as mathematics, psychology, music, science, etc. A learning object could be as small as a single step in the process of solving a problem. To match learning objects with learners' profiles at a specific point of a learning process, there have to be computable measures that sufficiently characterize learning objects and learners.

Knowledge is the core of a learning object. There are many aspects that need to be specified for a piece of knowledge. It is difficult to generate a universal list of all aspects because each tutoring system has its own specific considerations and the landscape is large. We name a few here that are common to many ITSs:

- **Prerequisites**. Pieces of knowledge are not acquired independently. A learner usually needs enough prerequisites in order to understand any given particular piece of knowledge. Learning objects need to be well organized so that the prerequisites of a given object can be identified. Knowledge space theory (Falmagne, Albert, Doble, Eppstein & Hu, 2013; Falmagne & doignon, 2011) offers a solid foundation for learning object organization.

- **Difficulty level**. Theoretically, if a learner has mastered all prerequisites, the learning object is "learnable" by the learner. However, a learning object with higher difficulty level may need a longer time to learn. If the difficulty level is too high, the object could be too challenging to some learners.

- **Level of detail**. A learning object with too much detail may look boring to high-ability learners whereas a learning object with too little detail may be too hard for a low-ability learner to comprehend. Adequate level of detail is important to maximize effectiveness, efficiency, and engagement.

Forms of knowledge representation in a learning object are also important in learning object selection. The knowledge may be represented by a mixture of texts, pictures, diagrams, audios, videos, simulations and conversations. Texts need to match the learner's reading comprehension level. Audios need to match listening comprehension level. Visual elements may pose different difficulty issues to different learners.

The pedagogical strategy embedded in a learning object is another important part. For example, in AutoTutor conversational tutoring, vicarious learning is appropriate for low-knowledge learners, tutoring is appropriate for medium-knowledge learners, and a teachable agent is appropriate for high-knowledge learners (Cai, Graesser, Forsyth, Burkett, Millis, Wallace, Halpern, D. & Butler, 2011; Cai, Graesser & Hu, 2015).

A learner's knowledge, ability, and personality can only be assessed from the interactions between the learner and the learning objects. Different types of interactions capture different characteristics about learners. Multiple-choice questions are perhaps the most popular type of interactions in learning systems. They are used to assess how much a learner knows about a piece of knowledge. The obvious advantage of using multiple-choice questions in learning systems is that the performance can be quickly and objectively scored as either correct or incorrect. With enough items, the percentage of correct answers approximates the learners' knowledge. There is no difficulty in automatically scoring answers to multiple-choice questions, but these questions can rarely capture information about the problem-solving process. In contrast, open-ended questions can capture much more information about learners' knowledge, ability, behavior, and personality (Dowell, 2017; Dowell, Graesser & Cai, 2016). However, the use of open-ended questions is still limited in most ITSs because it is hard to score their answers automatically.

In AutoTutor conversations, most natural language inputs from a learner are answers to open-ended questions. This chapter describes the detailed assessments at each conversation turn inside AutoTutor conversations. The turn by turn measures help AutoTutor give adequate immediate feedback to learners and adaptively select the best conversation paths. This chapter also discusses the use of a knowledge space model and a knowledge component model that are outside of the scope of AutoTutor's conversation mechanisms. The knowledge space model is used to organize AutoTutor knowledge objects, making sure that the knowledge objects cover a complete set of concepts representing a domain and that there is a path for a learner to learn new concepts one at a time. Knowledge space theory also helps select best learning path for a learner. The knowledge component model is used to trace learners' learning changes over time. It helps researchers deeply understand the knowledge objects, the learners, and the learning process. Natural language based assessment, knowledge space theory, and knowledge component model are recommended for the Generalized Intelligent Framework for Tutoring (GIFT).

AutoTutor Conversation

There are many conversational systems, such as ELIZA, Artificial Linguistic Internet Computer Entity (ALICE) Bot, Apple Siri, Amazon Echo, Google Home, etc. These systems provide interesting responses to a human's questions and comments. However, they do not provide deep and coherent conversations about a specific topic as does AutoTutor. A typical AutoTutor conversation starts with a main question selected by the system. The answer to the main question usually contains 3 to 10 sentences. The following is an example main question and its ideal answer in AutoTutor Physics (Graesser, Lu, Jackson, Mitchell, Ventura, Olney, et al., 2004):

- ***Main Question:*** *When a car without headrests on the seats is struck from behind, the passengers often suffer neck injuries. Why do passengers get neck injuries in this situation? Explain.*

- ***Answer to Main Question:*** *When a car is struck from behind the force of impact will cause a large forward acceleration in it. In order for the head to go along with the body it should have the same acceleration as the body. In the absence of a head support only the neck can exert this force on the head. In an attempt to produce the required large force, the neck gets stretched and may get injured damaging its muscles and ligaments.*

With the main question and an expert answer, a simple and minimally adaptive tutoring conversation can be set up in the following way:

1) AutoTutor asks the main question;
2) the learner gives an answer;
3) AutoTutor evaluates learner's answer;
4) AutoTutor gives a feedback; and
5) AutoTutor gives the ideal answer.

In terms of assessing a learner's knowledge and helping a learner learn, this simple conversation is probably just as good as a homework a student normally does.

Nevertheless, AutoTutor provides a much more complex conversation. The first step in constructing a complex conversation is to break an ideal answer into a set of *Expectations* (Nye, Graesser & Hu, 2014). An expectation in AutoTutor is a part of an ideal answer, usually about one sentence long. The entire set of the expectations for an ideal answer is semantically equivalent to the ideal answer. Breaking one ideal answer into multiple expectations makes it possible to more accurately assess a learner's knowledge and more deeply help learners learn. In the previous example, the ideal answer was broken into 4 expectations:

1) *When a car is struck from behind the force of impact will cause a large forward acceleration of the car.*
2) *In order for the person's head to go along with the person's body they should both have the same acceleration.*
3) *In the absence of a head support, the head is only accelerated in the neck.*
4) *In an attempt to produce the required large force, the neck gets stretched and may get injured damaging its muscles and ligaments.*

In this example, the 4 expectations are almost the same as the original 4 sentences in the ideal answer. However, in general, the expectations could be different from the ideal answer sentences. The sentences in an ideal answer need to be coherently connected, whereas the expectations are relatively independent knowledge pieces. A sentence in an ideal answer may be split into multiple expectations if it is too complex. That is, the number of expectations may be different from the number of sentences in an ideal answer.

AutoTutor assesses and helps a learner by asking questions around the expectations. There are two important types of questions AutoTutor often uses. One is called a "hint". A hint question is a question to which the answer is expected to be a clause, proposition or sentence. For example, the question "*How will the impact affect the car?*", the expected answer is "*The force of the impact will cause the car to experience a large forward acceleration.*" Another type of question is called a "prompt". A prompt question targets a word or phrase answer. For example, for the question "*The force of the impact will be directed _____?*", the answer is a word, "forward". Table 1 shows the hints and prompts for expectation 1 in the previous example.

AutoTutor helps learners to construct an answer that covers all expectations by asking hints and prompts. The conversation flow can be briefly described as follows:

1) Agent asks main question;
2) Student responds to main question;
3) Evaluate student's responses against the expectations associated with the main question and gives feedback;
4) If student's responses cover all expectation, go to 10); otherwise
5) Select an expectation that the student has not covered;

6) Agent asks a hint/prompt for the selected expectation;
7) Student responds to the hint/prompt;
8) Evaluate student's current response against current hint/prompt and gives feedback;
9) If student's responses cover the selected expectation or there is no more hint/prompt for the selected expectation, go to 4; otherwise go to 6;
10) Agent gives a closing summary remark and stops.

The maximum number of student turns in such a conversation is the number of questions (main question, hints and prompts). For the previous example, the domain experts prepared 11 hints and 17 prompts. That is, the conversation may go for as many as 29 student turns. Obviously, AutoTutor conversation provides more detailed help to students then does the simple conversation. Meanwhile, the conversation also provides more data for assessing a leaner's knowledge.

Table 1. Hints and prompts for the expectation "When a car is struck from behind the force of impact will cause a large forward acceleration of the car."

Type	Question	Answer
Hint	In terms of mechanics, what will happen when the car is struck from behind?	The car will experience a force that will accelerate the car.
Hint	How will the impact affect the car?	The force of the impact will cause the car to experience a large forward acceleration.
Hint	In what direction will the car experience a force?	The car will experience a force in the forward direction.
Prompt	The car suddenly accelerates because it has been _____?	because it has been struck.
Prompt	The force of the impact will result in the car's forward _____?	the car's forward acceleration.
Prompt	The force of the impact will be directed _____?	will be directed forward.
Prompt	The impact will result in a forward acceleration of the _____?	car and everything in it.

Knowledge Assessment Through AutoTutor Conversation

The lowest level of assessment in AutoTutor conversation is at each question step. AutoTutor evaluates how well a learner answers a question. Because an AutoTutor conversation starts with a main question, in each hint/prompt step, it is expected that a learner may not only answer the current question but also cover other expectations. There is an empirical question of whether this actually occurs. Our AutoTutor Physics data gives us a positive answer. Consider the following example:

- ***Hint Question:*** *Why will the head of the passenger need to be accelerated in the passenger's neck?*

- ***Expert Answer:*** *The head will be accelerated in the neck of the passenger because there is no headrest present.*

- ***Student Long Answer:*** *"It must have the same acceleration as the body to have the same position as the body and if there is no headrest the force causes the head to move with a large force and then when it snaps back toward the body, there is no headrest to stop the motion."*

The student answer in this example covers two expectations:

- *In order for the person head to go along with the person body they should both have the same acceleration.*
- *In the absence of a head support, the head is only accelerated in the neck.*

In the AutoTutor Physics data logs, the average length of expert answers to 114 hint questions is 11 words. However, in 4,941 hint answers we collected from students, about 20% the answers contained more than 12 words, 10% contained more than 17 words, and 5% more than 23 words (Figure 1.) Although the answers to prompt questions are expected to be 1–3 words, in the 2,643 student answers to prompt questions, there were still about 5% of the answers that contained more than 5 words and 1% more than 10 words. In short, a student answer may contain more information than the answer to the affiliated hint or prompt. Therefore, instead of comparing the semantic similarity between the expected answer of a hint/prompt and the student response, we need to consider the semantic "containment", namely, the extent to which the student answers has a subset of information that contains the expected answer.

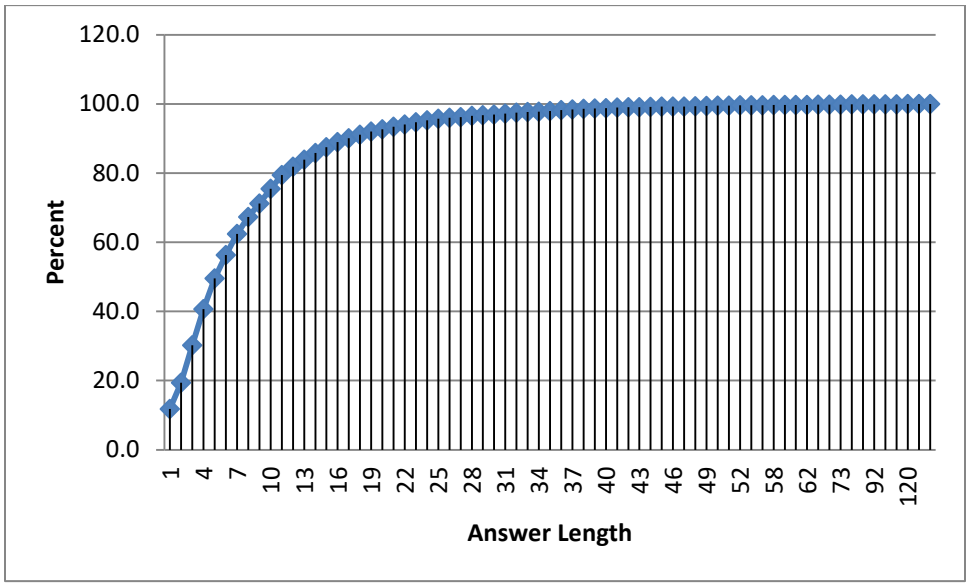


Figure 1. The cumulative percent of 4,941 student answers to 114 hint questions of different lengths.

Consider how this is worked out mathematically. Denote the main question as MQ, and the expectations as $E_1, E_2, \ldots, E_n$. Denote the questions asked in each step as $Q_1, Q_2, \ldots$, and the associated student responses as $R_1, R_2, \ldots$, where R_i is the response received after the question Q_i. At the i^{th} step, we need to evaluate the following:

1) How much does R_i contain the answer to Q_i?
2) How much does the "sum" of $R_1, R_2, \ldots R_i$ contain $E_1, E_2, \ldots, E_n$?
3) How much does the "sum" of $R_1, R_2, \ldots R_i$ contain the answer to MQ?

A semantic "contain" function that can evaluate how much a text A contains a text B can be defined in different ways. The simplest one is keyword matching, which can be defined as the proportion of keywords in the text B that appear in the text A. The keyword matching algorithm can be improved by replacing keywords with regular expressions. A regular expression defines a pattern of a character string. For example, "*\baccel*" can match any word staring with "*accel*"("*\b*" indicates "word boundary"). Using regular expressions can take care of minor misspelling and word derivatives. For example, if "*acceleration*" is a keyword, replacing it by regular expression "*\baccel*" will match misspelt word "*accelaration*" and derivatives of "*acceleration*" such as "*accelerate*", "*accelerates*", "*accelerating*", and "*accelerated*". The keyword matching can also be extended to allow matching synonyms. Each keyword can be expanded to a set of synonyms. For example, the word "*car"* may be expanded to a synonym set {"*car*", "*vehicle*", "*automobile*"}. A synonym set is matched if one of the words in the set is matched. The synonym sets can be further extended to regular expression sets.

This discussion leads to a problem of finding "synonyms". In fact, the so called "synonyms" are just alternative ways of expressing a given word or idea. It might be challenging for a domain expert to imagine all alternatives of a word or an idea at authoring time. Latent Semantic Analysis (LSA) can help (Foltz, Kintsch & Landauer, 1998; Foltz & Martin, 2008; Graesser, Penumatsa, Ventura, Cai & Hu, 2007; Gorman, Foltz, Kiekel, Martin & Cooke, 2003; Landauer, Foltz & Laham,1998; Olney, Louwerse, Mathews, Marineau, Hite-Mitchell & Graesser, 2003; Rus, Lintean, Graesser & McNamara, 2012). LSA represents words by vectors. The vector representations of words are obtained through singular value decomposition from a word-document matrix generated from a large corpus. It has been shown in many applications that the LSA vectors capture the semantics of words, in the sense that semantically similar words have similar vector representations. An LSA utility tool allows domain experts to choose synonyms from LSA nearest neighbors (i.e., the words with high LSA cosine values to the target words).

Based on LSA and RegEx, AutoTutor "contain" function can measure the quality of user's inputs as good as human does (Cai, et al., 2011). The contain function gives stepwise performance assessment on the coverage of each expectation, as well as the answer to the main question.

Immediate Feedback

The stepwise assessment helps AutoTutor give immediate feedback, which is helpful to learners (O'Neil, Chuang & Baker, 2010; Tausczik & Pennebaker, 2013). The assessment of the amount of an expected answer that is articulated by a student is used to give different types of feedback. If the amount is high, AutoTutor gives a positive feedback, such as "*Great job!*", "*Excellent!*", etc. If the amount is medium, the response is treated as a partial answer and AutoTutor may give a neutral feedback, such as, "*Not bad!*", "*Good try!*", "*OK.*", etc. If the amount is low, the response could be irrelevant to the answer, and AutoTutor may give irrelevance feedback, such as "*That doesn't answer the question*", "*That doesn't seem to be relevant*", etc.

One important category of feedback that is missing in this discussion is negative feedback, such as "*No*!", "*That is not right*!" Unfortunately, the semantic contain function can only tell the student if a response contains expected parts of an answer. It cannot tell whether an answer is correct or wrong. To avoid using any time-consuming reasoning algorithms, AutoTutor prepares and collects typical "bad answers" to questions. Thus, if a bad answer is matched, a negative feedback can be triggered. Domain experts may prepare possible bad answers. However, it is not possible for an expert to think of all possible ways student may

make mistakes. Therefore, collecting typical bad answers from real students is helpful. There is a difference between "bad answers" and "irrelevant responses". Bad answers are wrong answers but relevant to the topic. Fortunately, it is relatively easy to identify irrelevant answers from bad or good answers.

One persistent challenge has been that bad answers often share keywords with good answers. It would be misleading to give a negative feedback to a learner when the response is not wrong. To be safe, negative feedback is considered only when a bad answer is matched and no good answer is matched.

Behavior Assessment Through AutoTutor Conversation

The rich conversation data from AutoTutor make it possible to assess learners' behavior during the learning process. In each turn, it is possible to assess the following about the latest response:

- How much is new and relevant to our topic?
- How much is new but irrelevant to our topic?
- How much is old though relevant?
- How much is old and irrelevant?

A good learner is supposed to give new and relevant information in each turn. When a learner starts to give old and relevant information, it is probably an indicator that the learner has exhausted their knowledge about the topic. A learner who starts to give irrelevant response is probably tired. A learner who always gives irrelevant information is usually a "gamer", who is not really interested in learning. Hu et al. calls these four properties the learner's characteristic curves (LCCs; Hu, Cai, Han, Craig, Wang & Graesser, 2009; Hu, Nye, Gao, Huang, Xie & Shubeck, 2014).

The previous four properties cannot be easily computed through the keyword based contain functions. For example, the new information is the part of the latest response that is not contained in the previous turns. Using the keyword-based contain function implies that we need to identify the keywords in the learners' turns. Therefore, automatic keyword extraction is needed. Furthermore, if we want to use regular expressions, then automatic regular expression generation is needed, which is hard to implement.

There is an alternative way to compute these four measures using LSA (Hempelmann, Dufty, McCarthy, Graesser, Cai & McNamara, 2005). Learners' previous responses can be represented by vectors in LSA space by adding up (usually weighted) the word vectors in each of the responses. The vectors of previous responses $R_1, R_2, \dots R_{i-1}$ span a subspace. The vector of the i^{th}response R_i can then be decomposed into the sum of two components, one lies in the subspace and the other is perpendicular to the subspace. The length of the component in the subspace can be used as the amount of old information and the length of the perpendicular component can be used as the new information. Similarly, the expectations $E_1, E_2, \dots, E_n$ can form a subspace. The old information component and new information component can be further decomposed into old relevant, old irrelevant, new relevant and new irrelevant components. Over the turns, the length of the four components form the four LCCs.

Assessment Outside AutoTutor

In addition to drive the inside moves, the turn by turn assessment of AutoTutor conversations can be used outside AutoTutor. In this section, we briefly talk about two uses: 1) learning path selection in learning space theory (Falmagne, Albert, Doble, Eppstein & Hu, 2013; Falmagne & doignon, 2011) and 2) learning object and learning evaluation in knowledge component model (Koedinger, Baker, Cunningham, Skogsholm, Leber & Stamper, 2010).

In learning space theory, a learner's knowledge state about a topic is represented by a set of problem the learner could successfully solve. A complete set of problems should cover every part of a domain. All subsets of the problem set form the "points" of the learning space. The starting point is an empty set, representing the state of a learner who has no knowledge about the domain. The ending point is the whole problem set, representing a state of complete mastering of the domain. A learner learns by "walking" from the starting point to the ending point.

With enough problems, a learning space may provide many possible paths for a learner to walk through. The question is, how could the shortest path be selected? To provide the best path to a learner, the learning objects need to be well organized and adequate prerequisites should be identified. One way to do this is to present problem pairs to human experts and let experts make judgement on the order of the problems in problem pairs and thus infer the partial order (Falmagne & doignon, 2011). However, it is usually tedious and inaccurate. AutoTutor has detailed assessment on each problem. The final coverage to the answer of a main question can be used to judge the success (or failure) of an attempt. Therefore, AutoTutor assessment data can be used to infer the underline partial order of a problem set.

Knowledge component model associate problems with a relatively smaller set of knowledge components. Learners' performance on solving problems is mapped to the failures or successes in knowledge components. In solving a set of problems, a learner usually has multiple opportunities to work with each knowledge component. The success on a given component as a function of opportunities form a learning curve, which can be used to trace individual learner's learning as well as evaluating a learning object's difficulty level. An AutoTutor problem usually involves multiple knowledge components, which are distributed in the expectations. The assessment on expectation coverage provides success/failure information about knowledge components. Knowledge component model is implemented in Datashop and its new extension, LearnSphere (Veeramachaneni, Dernoncourt, Taylor, Pardos & O'Reilly, 2013; Stamper et al., 2016).

Recommendations and Future Research

GIFT provides a powerful platform for researchers to integrate ITSs. All ITSs have assessment modules in different forms. It is the assessment through natural language that provides rich information about learners' knowledge, ability, behavior, and personality. AutoTutor is currently an independent system in GIFT that is capable of assessing students through natural language; there are authoring tools to develop materials with AutoTutor. It is worthwhile for GIFT users to consider AutoTutor natural language conversation services that can be used in ITSs that would benefit from natural language interactions.

The natural language conversation metrics we discussed in this chapter will evolve to improved and new metrics. Continuous research and development on natural language conversation metrics is needed for GIFT. Rich assessment results are helpful for giving students immediate intelligent feedback and for selecting the best next step inside a learning object. Outside a specific intelligent system, accumulated assessment results are helpful in learning object selection. Learning space theory has a solid mathematical foundation and has been very successfully used in the Assessment and Learning in Knowledge Spaces (ALEKS) (Hoelzle & Bergman, 2000). We strongly recommend integrating knowledge components and knowledge space theory into GIFT as a domain knowledge organization.

Currently, AutoTutor data are saved in LearnSphere. Data saved in LearnSphere can be shared with many researchers and developers. We recommend saving other GIFT data to LearnSphere, so that more researchers and developers can benefit from data collected from GIFT. In return, new discoveries from researchers will facilitate further development of GIFT.

Acknowledgements

The research on was supported by the National Science Foundation (NSF) (DRK–12-0918409, DRK–12 1418288), the Institute of Education Sciences (IES) (R305C120001), US Army Research Laboratory (W911INF-12-2-0030), and the Office of Naval Research (N00014-12-C-0643; N00014-16-C-3027). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or Department of Defense. The Tutoring Research Group (TRG) is an interdisciplinary research team comprised of researchers from psychology, computer science, and other departments at University of Memphis (visit http://www.autotutor.org).

References

Cai, Z., Graesser, A.C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., Halpern, D. & Butler, H. (2011, November). Trialog in ARIES: User Input Assessment in an Intelligent Tutoring System. In W. Chen & S. Li (Eds.), Proceedings of the 3rd IEEE International Conference on Intelligent Computing and Intelligent Systems (pp.429–433). Guangzhou: IEEE Press.

Cai, Z., Graesser, A.C. & Hu, X. (2015). ASAT: AutoTutor script authoring tool. In. R. Sottilare, A.C. Graesser, X. Hu & K. Brawner (Eds.), Design Recommendations for Intelligent Tutoring Systems: Authoring Tools (Vol. 3) (pp.199–210). Orlando, FL: Army Research Laboratory.

Dowell, N.M. (2017). A computational linguistics analysis of learners' discourse in computer-mediated group learning environments. Dissertation, University of Memphis.

Dowell, N. M., Graesser, A. C. & Cai, Z. (2016). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. Journal of Learning Analytics, 3(3), 72–95.

Falmagne, J., Albert, D., Doble, C., Eppstein, D. & Hu, X. (2013). Knowledge spaces: Applications in education. Berlin-Heidelberg: Springer.

Falmagne, J. & doignon, J. (2011). Learning Spaces. Berlin: SpringerVerlag.

Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. Discourse Processes, 25, 285–307.

Foltz, P. W. & Martin, M. J. (2008). Automated communication analysis of teams. In E. Salas, G. F. Goodwin & S. Burke (Eds.), Team effectiveness in complex organisations and systems: Cross-disciplinary perspectives and approaches (pp. 411-431). New York, NY: Routledge.

Gorman, J. C., Foltz, P. W. Kiekel, P. A., Martin, M. A. & Cooke, N. J. (2003) Evaluation of Latent Semantic Analysis-based measures of communications content. In Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting.

Graesser, A.C. (2016). Conversations with AutoTutor help students learn. International Journal of Artificial Intelligence in Education, 26, 124–132.

Graesser, A.C., Forsyth, C.M. & Foltz, P. (2016). Assessing conversation quality, reasoning, and problem solving performance with computer agents. In B. Csapo, J. Funke, and A. Schleicher (Eds.), On the nature of problem solving: A look behind PISA 2012 problem solving assessment (pp. 275-297). Heidelberg, Germany: OECD Series.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A. M., et al. (2004). AutoTutor: A tutor with dialogue in natural language. Behavior Research Methods, Instruments & Computers, 36, 180–193.

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z. & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed initiative dialogue in natural language. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), Handbook of latent semantic analysis (pp. 243–262). Mahwah, NJ: Erlbaum.

Graesser, A. C., Person, N. K., Harter, D. & Tutoring Research Group. (2001). Teaching tactics and dialog in AutoTutor. International Journal of Artificial Intelligence in Education, 12(3), 257–279.

Hempelmann, C. F., Dufty, D., McCarthy, P., Graesser, A. C., Cai, Z. & McNamara, D. S. (2005). Using LSA to automatically identify givenness and newness of noun-phrases in written discourse. In B. Bara (Ed.), Proceedings of the 27th Annual Meeting of the Cognitive Science Society (pp. 941–946). Mahwah, NJ: Erlbaum

Hu, X., Cai, Z., Han, L., Craig, S. D., Wang, T. & Graesser, A. C. (2009, July). AutoTutor lite. In Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling (pp. 802–802). IOS Press.

Hu, X., Nye, B. D., Gao, C., Huang, X., Xie, J. & Shubeck, K. (2014). Semantic representation analysis: A general framework for individualized, domain-specific and context-sensitive semantic processing. In D.D. Schmorrow and C.M. Fidopiastis (eds.), Foundations of augmented cognition: Advancing human performance and decision-making through adaptive systems (pp. 35–46). Springer International Publishing.

Hoelzle, L. & Bergman B. (2000). Aleks. McGraw-Hill Pub. Co..

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B. & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero , S. Ventura & M. Pechenizkiy (Eds.), Handbook of educational data mining (Vol. 43). Boca Raton: CRC Press.

Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. Discourse Processes, 25, 259–284.

Nye, B.D., Graesser, A.C. & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. International Journal of Artificial Intelligence in Education, 24, 427–469. 2.

Olney, A., Louwerse, M., Mathews, E., Marineau, J., Hite-Mitchell, H. & Graesser, A. (2003). Utterance classification in AutoTutor. In J. Burstein & C. Leacock (Eds.), Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing. Philadelphia: Association for Computational Linguistics.

O'Neil, H. F., Chuang, S.H. and Baker, E.L. (2010). Computer-based feedback for computer-based collaborative problem-solving. In D. Ifenthaler, P. Pirnay-Dummer, N. M. Seel (Eds.), Computer-based diagnostics and systematic analysis of knowledge (pp. 261–279). New York, NY: Springer-Verlag.

Rus, V., Lintean, M., Graesser, A. C. & McNamara, D.S. (2012). Text-to-text similarity of statements. In P. McCarthy and C. Boonthum-Denecke (Eds.), Applied natural language processing: Identification, investigation, and resolution (pp. 110–121). Hershey, PA: IGI Global.

Stamper, J., Koedinger, K., Pavlik Jr., P. I., Rose, C., Liu, R., Eagle, M., . . . Veeramachaneni, K. (2016). Educational Data Analysis using LearnSphere Workshop. In J. Rowe & E. Snow (Eds.), Proceedings of the EDM 2016 Workshops and Tutorials co-located with the 9th International Conference on Educational Data Mining. Raleigh, NC. Workshop.

Sottilare, R., Graesser, A., Hu, X. & Goldberg, B. (Eds). (2014). Design Recommendations for Intelligent Tutoring Systems: Volume 2 - Instructional Management. Orlando, FL: US Army Research Laboratory.

Tausczik, Y. R. & Pennebaker, J. W. (2013). Improving teamwork using real-time language feedback. Proceedings of Human Factors in Computing Systems (CHI), 459–468.

Veeramachaneni, K., Dernoncourt, F., Taylor, C., Pardos, Z. & O'Reilly, U.-M. (2013). MOOCdb: Developing data standards for MOOC data science. In E. Walker & C.-K. Looi (Eds.), AIED 2013 Workshops Proceedings Volume (pp. 17–24).

Zapata-Rivera, D., Jackson, G.T. & Katz, I. (2015). Authoring conversation-based assessment scenarios. In R. Sottilare, X. Hu, A. Graesser & K. Brawner (Eds.), Design Recommendations for Adaptive Intelligent Tutoring Systems. (Vol.3) (pp.169–178). Orlando, FL: US Army Research Laboratory.

CHAPTER 28 – Assessment of Individual Learner Performance in Psychomotor Domains

Jong W. Kim[1], Robert A. Sottilare[1], Gregory Goodwin[1], and Xiangen Hu[2]
US Army Research Laboratory[1], University of Memphis[2]

Introduction

The goal of training is to support the learner to achieve the learning objectives, i.e., reducing the task completion time, reducing errors, and increasing accuracy. Performance characterized by such speed and accuracy can be evaluated by interpreting the learner behavior – that is, as one of the classic models, Fitts' law provides a way to assess a regularity of a simple psychomotor task, such as tapping or pointing (Fitts, 1954). Based on the Fitts' model, researchers have investigated the coordination of physical human movement and information processing. The model has been used and expanded to predict the time to position and move around a mouse cursor in the human-computer interaction area (e.g., MacKenzie, 1992).

In this line of research, another popular classic model to assess the learner behavior is an engineering model of keystrokes (i.e., KLM-GOMS), which provides a prediction time of typing skill of an expert (Card, Moran & Newell, 1983). The model includes a physical operator and a mental operator so that it can describe the coordination of cognitive and physical functions within the learner to perform a task. The task used in that model is an interaction with a keyboard and mouse including keystrokes, mouse movements, and mouse clicks. These quantitative models have brought an important insight to assess psychomotor tasks. However, KLM-GOMS does not account for learning and performance change. It only predicts an expert's performance. As a recent advance, a cognitive architecture (e.g., Adaptive Control of Thought—Rational [ACT-R]) appeared to start answering such questions of learning – i.e., how to model psychomotor performance of keystrokes and how to predict human learning and performance, which has been also widely validated by cognitive and brain-imaging experiments (Anderson, 2007; Anderson et al., 2004).

However, assessment is mostly limited to the desktop environment (e.g., interacting with a computer). Such models are unable to explain the process of achieving the learning objectives and predict performance from a novice to an expert beyond the desktop environment. In terms of domain definition and complexity, psychomotor tasks can be classified into low and high complexity and ill- and well-defined domain (Sinatra & Sottilare, 2016). The aforementioned psychomotor tasks including tapping, pointing, and keystrokes can be considered as a task with low complexity in a well-defined domain. It is, thus, necessary to provide an advanced assessment method to better understand complex psychomotor tasks in sports, military, and medical domains, i.e., minimally invasive surgery skills, cardiac life support (e.g., tracheal intubation), throwing a ball, driving a car, marksmanship, dancing, golfing, and archery.

Behavioral assessment data are usually related to speed and accuracy, and are useful to identify psychomotor ability of an individual to assess psychomotor performance. Psychomotor ability indicates processing speed and accuracy observed in human performance, which is usually represented in the speed of responses to stimuli with little or no demands on cognitive processing and is mostly independent of information processing (e.g., Ackerman, 1988; Dosher, 1976; Wickelgren, 1981). Typically, psychomotor ability can be measured through simple reaction time (e.g., Seibel, 1963), the rate of movements (e.g., Fleishman, 1954), spatial orientation (Fleishman & Hempel Jr., 1955), two-hand coordination (e.g., Fleishman & Rich, 1963), and tapping (e.g., Fitts, 1954). Visuospatial and perceptual abilities may be also useful for assessment. That is, as the result of such assessments, the selection process of trainees for higher surgical training can help to determine which candidates are best suited to a certain surgery.

These prototypical measures would indirectly play a fundamental role to assess a complex psychomotor task (e.g., minimally invasive surgery, land navigation, etc.). A study shows that there is difference in the time to complete the task and errors by the skill level from a novice to an experienced doctor (Gallagher et al., 2001). However, we might encounter insufficiency of such assessment when we try to answer 1) how do we help the learner to develop expertise of surgical skill precision in terms of training efficiency?, 2) why is a professional golfer's putting performance in a time-pressured and stressful competition environment different from the previous training performance; 3) is a critical psychomotor skill, that might be rarely used and decay over time, ready to respond an emergency?, and 4) what is the training strategy to deal with pre- and post-stress from psychomotor related performance? It is, therefore, necessary to pursue investigating such questions to better achieve the learning objectives of training. It is necessary to advance our understanding of the coordination of cognitive and physical activities with physiological changes (e.g., respiratory and heart rates), and generalize such factors with consideration of the taxonomy of tutoring domains (Sinatra & Sottilare, 2016), which helps us to devise an advanced intelligent training system.

This approach can provide a much broader understanding of psychomotor performance and its assessment. For example, a golfer's putting performance can be assessed and understood by attentional resources during the physical activity of hitting the ball (the coordination of cognitive and physical functions), and it is also possible to unobtrusively measure the variability of physiological factors (i.e., heart rate variability and respiratory activity) to assess psychomotor performance (Lagos et al., 2011; Neumann & Thomas, 2009, 2011). A golfer would hold breath briefly when hitting the ball. A golfer's performance would be affected by fatigue or sleep deprivation as well. Data related to motion capture of human performance in VR can be also used to assess psychomotor tasks as well. Thus, the approach from the physical, cognitive, and physiological standpoints can make more sense of psychomotor performance data. In the next section, we describe psychomotor task performance with a functional relationship.

Psychomotor Skill Learning and Performance

We describe the process of the learner behavior by practice as a tool for psychomotor task assessment. The process toward achieving the learning objectives is generally summarized as a power law of learning (or practice). That is, learning behavior generally follows a regularity known as a power law of practice (e.g., Card, English & Burr, 1978; Delaney et al., 1998; Newell & Rosenbloom, 1981; Seibel, 1963). In addition, forgetting behavior is also known as a regularity of a power function or an exponential function (e.g., Anderson, Fincham & Douglass, 1999; Pavlik & Anderson, 2005; Rubin & Wenzel, 1996). An understanding of learning (and forgetting) behavior in psychomotor tasks is an important way to assess the learner behavior and provide adaptive instruction. For example, a learning (forgetting) curve of a psychomotor skill (i.e., hitting a golf ball) might be different from the one of a cognitive skill (i.e., solving a math problem). A study reports microgenetic analysis of subtasks, arguing that there exists different learning curves by different subtask skills (Kim & Ritter, 2016).

A Theory for Psychomotor Skill Learning

Based on the aforementioned Fitts' experiment (i.e., a simple motor performance tapping, and pin transfer)(Fitts, 1954), he asserted that there are stages of psychomotor skill learning (early, intermediate, and late stages), which is primarily continuous and is a kind of gradual shifts in the skill structure (Fitts, 1964). Starting from the Fitts' argument, there has been a consensus understanding toward a three-stage of skill learning (see Anderson, 1982; Rasmussen, 1986; VanLehn, 1996). This learning behavior is an important feature to understand and assess psychomotor performance.

This consensus understanding is also computationally explained by a cognitive architecture ACT-R (Anderson, 2007; Anderson et al., 2004). That is, the first stage is for acquiring declarative knowledge to perform the task, the second stage is for consolidating the acquired knowledge, and the third stage is for tuning the knowledge toward overlearning. Based on this consensus understanding, it is suggested that both skill learning and retention should be taken into consideration in this three-stage process (Kim & Ritter, 2015). Figure 1 shows the three stages of learning and retention, providing important insights about how learning and forgetting would be different at each stage. The main continuous line indicates continuous practice, which follows a regularity known as a power law of practice. Dashed lines indicate periods of no practice, with solid lines showing later training.

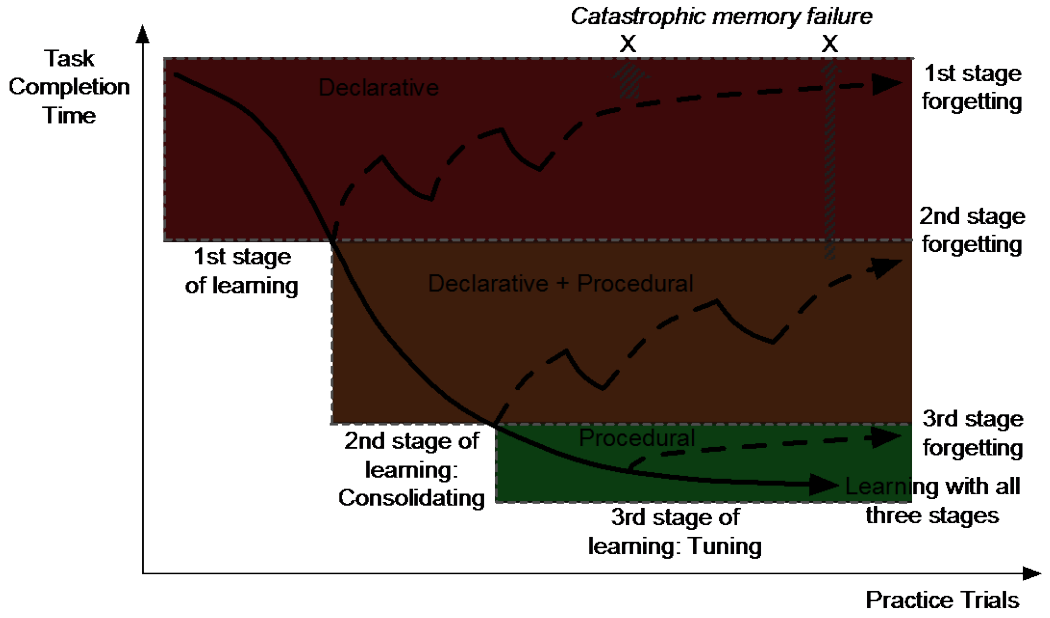


Figure 1. A theory of skill learning and retention.

The First Stage: Declarative. In the first stage, skill acquisition occurs and simple training focused on skill acquisition may be adequate. Task knowledge in declarative memory degrades with lack of use, perhaps catastrophically as indicated by X's in Figure 1, leading to inability to perform the task. With lack of use, the strength of declarative memory declines. Decreased memory strength leads to response time increasing and accuracy decreasing. In addition, the ACT-R theory explains that increases in working memory load leads to decrements of retrieval performance from memory based on the activation mechanism (see Anderson, Reder & Lebiere, 1996). Increase in working memory load can impair performance with this level of knowledge.

The Second Stage: Associative. In the second stage, task knowledge is represented with a mix of declarative and procedural memory. With lack of use, declarative knowledge can be degraded, leading to missed steps. Procedural memory, on the other hand, is basically immune to decay. In the first and second stage, catastrophic memory failure can occur because declarative knowledge is not fully activated. In this mixed stage, training should be provided to keep declarative knowledge active and also support further proceduralization.

The Third Stage: Procedural. In the third stage, task knowledge is available in both declarative and procedural forms, but procedural knowledge predominantly drives performance. Practice will compile knowledge into procedural knowledge. We refer to this type of task knowledge as a proceduralized skill. With lack of use, declarative knowledge may be degraded. Nevertheless, learners can still perform the task

– if all the knowledge is proceduralized and thus not forgotten with time. Less well-known skills that are infrequently used, like recovery from unusual errors, may be degraded. This type of skill would require knowledge retrieval from declarative memory unless task knowledge is proceduralized.

This learning and retention theory, based on ACT-R (Anderson, 2007; Anderson et al., 2004), describes human learning as a three-stage process with an emphasis on a distinctive classification of the types of task knowledge: declarative and procedural. Declarative knowledge is represented as a propositional network of facts consisting of chunks. Procedural knowledge, represented as production rules, refers to knowledge that is displayed in behavior such as steps and sequences of how to do a task. This knowledge classification would be useful to construct the ontological representation of task knowledge.

As an example, supposed that learning a typing skill. Individuals first memorize the layout of the keyboard; this is the first stage, declarative stage; like where is the letter, "a"?, the letter, "a", is next to the letter, "s". Practice enables individuals to memorize the layout and type faster. Over time, practicing typing skill, declaratively learned knowledge, leads to procedural knowledge – that is, rather than retrieving the location of the letter "a", individuals just imagine typing the letter and sees where their finger goes. This example illustrates how individuals use both declarative and procedural knowledge in memory, and execute such a perceptual-motor skill into an action.

Attentional Resources and Pressure

To provide a successful training regimen, it is crucial to understand what mechanisms are responsible for skill learning and performance under a real-world situation (e.g., under time pressure or other stress factors), and what functional relationships between cognitive (attentional resources) and physical functions in terms of the skill level represented in Figure 1. We believe such considerations will bridge the gap between training in simulated environments and performance in real-world situations.

As the task skill is practiced, it can be assumed that attentional resources are reduced in executing the task skill, but spare working memory capacity may be required to transition between stages (e.g., Sweller, 1988). In the early stages, more attentional resources are required to execute the skill (i.e., high information access cost). On the other hand, in the later stage (i.e., the third stage), the task skill is executed without excessive effort as related to attentional resources (i.e., low information access cost).

What governs this behavioral change and execution of a physical activity of the trainee? Is it because attentional focus is shifted to task-irrelevant cues (e.g., Easterbrook, 1959; Wine, 1971)? Is it because there is increase in attention that is being paid to step-by-step execution of the task skill set rather than the proceduralized skill set in the later stage of learning (e.g., Baumeister, 1984; Lewis & Linder, 1997)? These arguments are grounded in two competing theories among researchers: distraction theories and explicit monitoring theories. As mentioned earlier, the three stages, where performance capacity changes occur, require different attentional resources. That is, in the earlier stage, if the task skill execution depends on retrieval of memory items in declarative memory, a stress factor would create the potential distraction to shift attentional focus to task-irrelevant cues such as worries, which is based on a process known as distraction theories. In the early stage, performance change by distraction theories can be represented as the strength (or probability) of retrieval in declarative memory. This is formalized as the activation mechanism in the subsymbolic level of the ACT-R architecture. In ACT-R, the base-level activation is dependent on how often (frequency) and how recently (recency) a chunk is used. Whenever a chunk is presented, the base-level activation increases, and then decreases as a power function of the time. The time to complete a task (e.g., latency) appears to decrease as a power function of the trial numbers of practice (see Anderson, Fincham & Douglass, 1999).

Another relevant theory applies to explicit monitoring of task skill execution. In the middle and later stages, task skills are *proceduralized*, indicating execution of task skill is largely unattended without the service of working memory, like expertise in typing (Posner, 1973) and baseball batting (Gray, 2004). This behavior can be represented as the knowledge compilation mechanism indicating two production rules are compiled to one, which leads to a faster process time (Anderson, 1982, 1987). In this explicit monitoring theory, a stress factor raises anxiety about performing correctly, which causes the reversion of attentional focus to step-by-step control of skill processes (e.g., Baumeister, 1984; Lewis & Linder, 1997). This theory can provide an explanatory account for performance failure in the later stage.

Beilock et al. (2004; 2001) pointed out that the aforementioned theories have been seemingly considered to be mutually exclusive but should, in fact, be considered to be complementary. This complementary understanding is possible when we consider the three stages of the learner behavioral change, shown in Figure 1. That is, under the distraction theory, task skills, in the early stage, rely more on the retrieval process of an individual declarative memory item. In this stage of skill learning and performance, individuals depend almost exclusively on declarative memory elements to perform the task; this first stage is cognitively intensive and slow; information access cost is high. Information access cost is higher than in the later stage, and, task skills, in the later stages, rely on production rule learning (i.e., proceduralization). In the second stage, individuals begin to rely more on procedural memory elements but still rely on the declarative knowledge of the keyboard layout. Finally individuals progressively evolve into experts; they shift entirely (almost entirely) to using procedural memory; information access cost is low.

Early experimental work by Posner (1973) showed that procedural memory is more robust. In Posner's experiment, skilled typists were asked to label a diagram of a standard keyboard. He reported that the skilled typists had difficulty in recalling a visual location of a letter from the standard keyboard (declarative memory), whereas the typists could type the letters in a few seconds without errors. This example supports declarative knowledge of visual location can be degraded while procedural knowledge can remain robust against decay, suggesting that long-term retention can be possible when declarative knowledge turns into procedural knowledge. Theoretically, procedural form of knowledge is, therefore, the knowledge that we want in a real situation, which is more robust against stress and time pressure.

As seen in the typing example, individuals acquire knowledge and skills, and store them into memory. Training and practice move the acquired knowledge and skill to a certain stage. Theoretically, performance change from novice to expert can be explained by differential information access cost with regard to the three stages. Thus, infrequent use of knowledge (e.g., infrequent training of emergency response skill) can cause performance decrements because information access cost is high, leading to forgetting or catastrophic memory failure.

Breathing Properly or Choking Badly Under Pressure

We sometimes observe that a professional athlete performs more poorly than expected. In cognitive psychology, this phenomenon is called *choking under pressure* (Beckmann, Gröpel & Ehrlenspiel, 2013; DeCaro et al., 2011; Gray, 2004; Lewis & Linder, 1997). For example, in high stressful situations, a golfer who is endeavoring to make the cut for the PGA tour would perform more poorly than their skill level. The aforementioned distraction and explicit monitoring theories can partly account for the phenomenon, but it is curious that how the physiological factor is interrelated with attentional resources in terms of the three stages of learning (and retention). Physiological factors (e.g., heart rate [HR] variability [HRV] and respiratory activity) should be considered as other sources to assess psychomotor performance.

Newmann and Thomas (2009, 2011) investigated measures of cardiac and respiratory activities when individuals at different levels of skill developments during the a golf putting task. Compared to a novice golfer,

the expert golfers showed a pronounced phasic deceleration in HR immediately prior to the putt, and greater HRV in the very low frequency band, and a greater tendency to show a respiratory pattern of exhaling immediately prior to the putt (Neumann & Thomas, 2009). In a follow-up investigation of Neumann and Thomas, participants performed the putting task to measure both cardiac and respiratory activity under with or without attentional focus instructions (Neumann & Thomas, 2011). The results show that the experienced and elite golfers showed better performance and reduced HR, greater HRV, pronounced HR deceleration prior to the putt, and a greater tendency to exhale prior to the putt, compared to novice golfers. This study shows a relationship between psychomotor performance, physiological factors, and the skill level.

It is reported that a range of HRs are related to psychomotor skill performance – i.e., around 115 beats per minute (bpm), fine motor skills are beginning to deteriorate, and complex psychomotor skills are degraded around 145 bpm, and gross motor skills (e.g., running) start to break down above 175 bpm (Grossman & Christensen, 2008, pp. 31). As a training regimen, a tactical breathing method is used to address psychomotor performance under pressure (e.g., Grossman & Christensen, 2008), and, it is even reported that a breathing technique can lower blood pressure as well (Grossman et al., 2001). Furthermore, there is a report that psychological performance training including tactical breathing help to manage stress; i.e., tactical breathing and mental imagery can mitigate negative effects of stress for police officers (e.g., Page et al., 2016), and stress management training with tactical breathing is effective in reducing stress in soldiers (e.g., Bouchard et al., 2012). As a technique to delink memory from a physiological arousal, soldiers are trained to do tactical breathing to lower HRs.

In the study by Bouchard et al. (2012), one group of participants received a usual training with no session of supervised practice, and the other group received biofeedback informing the participants' current level of arousal under an immersive 3-D simulation and training environment. Participants in the latter group trained tactical breathing as a tool to deal with stress. The main measure to assess the level of stress was the concentration of salivary. This study confirms the use of stress management training including tactical breathing is effective in reducing stress.

Data for Psychomotor Performance Assessment

Assessing psychomotor tasks, which generally involves the coordination of cognitive and physical functions, pose significant challenges in capturing behavioral data (e.g., detecting physical movement) and assessing the alignment of those behaviors with a model of expert behavior. Thus, it is necessary to unobtrusively sense physical movements and physiological measures to determine the state of the individual learner and how it varies from expert models to assess the rate of progress toward the learning objectives and the development of psychomotor skills. In this section, we describe recent advances in a less obtrusive way to sense physical activity and physiological changes.

Sensing Physical Activity in the Wild

Intelligent tutoring systems (ITSs) have shown greater impacts: to note a few, in procedural troubleshooting tasks (e.g., Lesgold et al., 1992) and mathematics problem-solving tasks (e.g., Anderson, Boyle & Reiser, 1985). However, the tasks are not usually related to psychomotor tasks. It is, therefore, necessary to support a training paradigm beyond the desktop environment to support the psychomotor task training (e.g., using a Google glass to present content from an intelligent tutoring system, Sottilare, 2015; Sottilare & LaViola, 2015). That is, a lecture type training based on PowerPoint slides would not be enough. One page summary of the putting instruction would be simple, but achieving expertise is another story. For example, a novice golfer needs to acquire a set of task knowledge to perform a putting task. A golfer sequentially executes a series of mechanical actions: 1) position the ball, 2) align shoulders, hips, knees, and feet, 3) check postures of grip, standing, arms, hands, and head, 4) check weight distribution, 5) stroke, and 6) keep appropriate

postures after stroke. To increase accuracy, a golfer would need to control breathing as well (e.g., tactical breathing) at the same time.

Mobile phones are rapidly adopted and have been a medium to develop applications for an individual, a group, and a community scale sensing – i.e., the sensors include accelerometer, digital compass, gyroscope, GPS, microphone, camera, Wi-Fi, and Bluetooth (Lane et al., 2010). This smartphone-based sensing system can be a good medium to provide a user's current activity and behavior, which is useful to assess the learner's skill level and performance (e.g., Nguyen et al., 2015; Rai et al., 2012).

For example, GPS, to locate where the phone is, can be used to a psychomotor task of land navigation. The accelerometer and gyroscope data are capable of characterizing physical movements, i.e., distinct patterns from the accelerometer data can be exploited to recognize different physical activities such as putting, swing, running, walking, and standing. A microphone and camera can also be used to sense ambient sound whether it is stressful or not. More sensors can be incorporated into the functionality of the smartphone. For example, a barometer, which measures atmospheric pressure, can be used with the accelerometer to identify whether the user performs physical activities of walking or climbing (ascending or descending).

There are several frameworks that use mobile sensing data and provide user activity in a large-scale time series analysis. One of such platform is MobiSens (Wu, Zhu & Zhang, 2013) that can recognize a user's various activities such as sitting, running, and walking. It also affords ground reporting information in battle fields, i.e., firefighting scenes and disaster responses with the user's current position and activity. Sensing-Kit (Katevas, Haddadi & Tokarchuk, 2014) is another mobile sensing framework that is open source and supports both iOS and Android-based smartphones. EmotionSense and Funf are more oriented to gather a user's emotion and mood. Table 1 summarizes the current frameworks for mobile sensing of a large-scale user activity.

Table 1. A list of mobile sensing frameworks

	Data-Capturing Features	Data-Extraction Format	Platform Support
MobiSens	Accelerometer, magnetometer, GPS, light, sound recording, environment temperature, Bluetooth, power consumption	n/a	Android
SensingKit	Motions: accelerometer, gyroscope, and magnetometer Location: GPS Proximity: Bluetooth Smart data	JSON, CSV	iOS, Android
EmotionSense	Self-reported mood, accelerometer, sociability (calling and texting), GPS for location	JSON	Android
Funf	GPS for location, accelerometer (for sleeping and waking patterns), phone usage, temperature	SQLite file	Android

Sensing Physiological Changes

Biofeedback, which refers to a technique to quantitatively and unobtrusively measure bodily functions including heart rate, brain waves, skin temperature, blood pressure, muscle tension, and respiratory activity, can be useful to assess psychomotor performance as well. Based on the basic mobile sensing capability, it

is possible to measure cardiovascular and respiratory activities with an additional attachment (e.g., a photoplethysmogram-based sensor that gathers volumetric measures).

Respirator sensors, in general, use two types of techniques: 1) impedance pneumography and 2) inductive plethysmography. It is reported that the latter technique is a newer approach and provides a higher degree of accuracy (e.g., Zhang et al., 2010). When it comes to the measurement of HR, earphones can be used-sensor (i.e., photoplethysmographic sensors) is integrated into earphones to collect HR measurements during psychomotor performance (e.g., Poh et al., 2009). The data from these sensors can be transmitted to a mobile device through Bluetooth that is a default functionality these days.

A Novel Training Paradigm for Psychomotor Tasks

Resources are invested to train soldiers, which is primarily intended to provide competency, preparedness, and readiness to effectively address a certain situation. To achieve that end, researchers investigate, develop, and deploy intelligent systems for training, such as an adaptive tutoring system in an attempt to help soldiers practice knowledge and skills. This ITS seeks to provide advanced authoring and maintaining for individualized and self-paced training regimens. One of such systems is the Generalized Intelligent Framework for Tutoring (GIFT), implemented and maintained by US Army. Such systems have been proved to improve training effectiveness and efficiency – an ITS is to help soldiers to get sufficiently trained and be ready for a certain mission in a special operational environment. The current manifestation of ITSs is, however, mostly restricted to a desktop environment. A real operational environment would be usually constrained by time or other stressors (e.g., time pressure, fear, and worry), which would limit both physical and cognitive performance.

Soldiers need to train a wide range of psychomotor skills (i.e., coordination of both physical and cognitive performance), and execute them under time-critical and stressful situations. It is, particularly, necessary to focus on precision-required psychomotor skill training (e.g., shooting a moving target). From a theoretical perspective, it is important to move the skill set to the later stage where the robust knowledge and skill structure is to be formed; this skill set can be more resistant against stressors (e.g., time pressure). To efficiently move the psychomotor skill set into the later stage, it is necessary to advance an intelligent tutoring framework using adaptive strategies by assessing the soldier's learning and performance.

To strengthen the current GIFT capability, we discussed theoretical accounts about assessment of learning and performance in this chapter. One of the identified needs is to support psychomotor skill training beyond the desktop environment so that it can minimize the performance gap between the conventional instructional environment and the real operational environment (Sottilare, 2015; Sottilare & LaViola, 2015). Thus, it is necessary to identify technical needs and research questions to minimize such gaps, and improve the current GIFT capability toward a better support for adaptive instructional strategies of the psychomotor task training. It is also necessary to identify the mechanisms of performance change (i.e., the skill level from a novice to an expert) toward acquisition of extreme expertise by looking at both attentional and physiological properties (i.e., focus of attention, breathing, heart rate variability) since it would be useful to providing a science-based adaptive instruction and feedback of a tactical breathing technique by extending the current GIFT framework.

GIFT can be used to obtain a real-time feedback (assessment of psychomotor performance) from the mobile sensing device. The GIFT authoring tools (GAT) also provide a structure to develop adaptive instruction for psychomotor tasks (Goldberg, 2016; Sottilare, 2015; Sottilare & LaViola, 2015). As an extension of GIFT toward "adaptive tutoring on the run", it is worth exploring the capability of the smartphone-based activity monitor in a training environment, such as 1) storing and analyzing the logged trainee performance, 2) authoring individualized tutoring contents, 3) implementing and evaluating the effectiveness of skill

transfer, and 4) iterating the process of personalized instruction. A cognitive modeling approach should be also considered as well since this approach can support predictive performance that can be used by the tutor.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*(3), 288–318.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369–406.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94*(2), 192–210.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C. & Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 111*(4), 1036–1060.

Anderson, J. R., Boyle, C. F. & Reiser, B. J. (1985). Intelligent tutoring systems. *Science, 228*(4658), 456–462.

Anderson, J. R., Fincham, J. M. & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(5), 1120–1136.

Anderson, J. R., Reder, L. M. & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology, 30*(3), 221–256.

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology, 46*(3), 610–620.

Beckmann, J., Gröpel, P. & Ehrlenspiel, F. (2013). Preventing motor skill failure through hemisphere-specific priming: Cases from choking under pressure. *Journal of Experimental Psychology: General, 142*(3), 679–691.

Beilock, S. L., Bertenthal, B. I., McCoy, A. M. & Carr, T. H. (2004). Haste does not always make waste: Expertise, direction of attention, and speed versus accuracy in performing sensorimotor skills. *Psychonomic Bulletin & Review, 11*(2), 373–379.

Beilock, S. L. & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General, 130*(4), 701–725.

Bouchard, S., Bernier, F., Boivin, É., Morin, B. & Robillard, G. (2012). Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers. *PLoS ONE, 7*(4), e36169.

Card, S. K., English, W. K. & Burr, B. J. (1978). Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT. *Ergonomics, 21*(8), 601–613.

Card, S. K., Moran, T. P. & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum.

DeCaro, M. S., Thomas, R. D., Albert, N. B. & Beilock, S. L. (2011). Choking under pressure: multiple routes to skill failure. *Journal of Experimental Psychology: General, 140*(3), 390–406.

Delaney, P. F., Reder, L. M., Staszewski, J. J. & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9*(1), 1–7.

Dosher, B. A. (1976). The retrieval of sentences from memory: A speed-accuracy study. *Cognitive Psychology, 8*(3), 291–310.

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review, 66*(3), 183–201.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology, 47*(6), 381–391.

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 243–285). New York: Academic Press.

Fleishman, E. A. (1954). Dimensional analysis of psychomotor abilities. *Journal of Experimental Psychology, 48*(6), 437–454.

Fleishman, E. A. & Hempel Jr., W. E. (1955). The relation between abilities and improvement with practice in a visual discrimination reaction task. *Journal of Experimental Psychology, 49*(5), 301–312.

Fleishman, E. A. & Rich, S. (1963). Role of kinesthetic and spatial-visual abilities in perceptual-motor learning. *Journal of Experimental Psychology, 66*(1), 6–11.

Gallagher, A. G., Richie, K., McClure, N. & McGuigan, J. (2001). Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. *World Journal of Surgery, 25*(11), 1478–1483.

Goldberg, B. (2016). Intelligent tutoring gets physical: Coaching the physical learner by modeling the physical world. In *Proceedings of the 10th International Conference on Augmented Cognition-HCII2016* (pp. 13–22). Toronto, Canada: Springer.

Gray, R. (2004). Attending to the execution of a complex sensorimotor skill: expertise differences, choking, and slumps. *Journal of Experimental Psychology: Applied, 10*(1), 42–54.

Grossman, D. & Christensen, L. W. (2008). *On combat: The psychology and physiology of deadly conflict in war and in peace* (3rd ed.): Warrior Science Publications.

Grossman, E., Grossman, A., Schein, M., Zimlichman, R. & Gavish, B. (2001). Breathing-control lowers blood pressure. *Journal of Human Hypertension, 15*(4), 263–269.

Katevas, K., Haddadi, H. & Tokarchuk, L. (2014). Poster: Sensingkit: A multi-platform mobile sensing framework for large-scale experiments. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking* (pp. 375–378). ACM.

Kim, J. W. & Ritter, F. E. (2015). Learning, forgetting, and relearning for keystroke- and mouse-driven tasks: Relearning is important. *Human-Computer Interaction, 30*(1), 1–33.

Kim, J. W. & Ritter, F. E. (2016). Microgenetic analysis of learning a task: Its implications to cognitive modeling. In F. E. Ritter & D. Reitter (Eds.), *Proceedings of the 14th International Conference on Cognitive Modeling* (pp. 21–26). University Park, PA: Penn State.

Lagos, L., Vaschillo, E., Vaschillo, B., Lehrer, P., Bates, M. & Pandina, R. (2011). Virtual reality-assisted heart rate variability biofeedback as a strategy to improve golf performance: A case study. *Biofeedback, 39*(1), 15-20.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T. & Campbell, A. T. (2010). A survey of mobile phone sensing. *IEEE Communications magazine, 48*(9), 140–150.

Lesgold, A. M., Lajoie, S. P., Bunzon, M. & Eggan, E. (1992). SHERLOCK: A coached practice environment for an electronics troubleshooting job. In J. Larkin, R. Chabay & C. Scheftic (Eds.), *Computer assisted instruction and intelligent tutoring systems: Establishing communication and collaboration*. Hillsdale, NJ: Lawrence Erbaum.

Lewis, B. P. & Linder, D. E. (1997). Thinking about choking? Attentional process and paradoxical performance. *Personality and Social Psychology Bulletin, 23*, 937–944.

MacKenzie, I. S. (1992). Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction, 7*, 91–139.

Neumann, D. L. & Thomas, P. R. (2009). The relationship between skill level and patterns in cardiac and respiratory activity during golf putting. *International Journal of Psychophysiology, 72*(3), 276–282.

Neumann, D. L. & Thomas, P. R. (2011). Cardiac and respiratory activity and golf putting performance under attentional focus instructions. *Psychology of Sport and Exercise, 12*(4), 451–459.

Newell, A. & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Lawrence Erlbaum.

Nguyen, L. T., Tague, P., Zeng, M. & Zhang, J. (2015). Superad: Supervised activity discovery. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 1463–1472). Osaka, Japan: ACM.

Page, J. W., Asken, M. J., Zwemer, C. F. & Guido, M. (2016). Brief mental skills training improves memory and performance in high stress police cadet training. *Journal of Police and Criminal Psychology, 31*, 122-126.

Pavlik, P. I. & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science, 29*, 559–586.

Poh, M.-Z., Kim, K., Goessling, A. D., Swenson, N. C. & Picard, R. W. (2009). Heartphones: Sensor earphones and mobile application for non-obtrusive health monitoring. In *International Symposium on Wearable Computers* (pp. 153–154). IEEE Computer Society.

Posner, M. I. (1973). *Cognition: An introduction*. Glenview, IL: Scott, Foresman and Company.

Rai, A., Yan, Z., Chakraborty, D., Wijaya, T. K. & Aberer, K. (2012). Mining complex activities in the wild via a single smartphone accelerometer. In *Proceedings of the Sixth International Workshop on Knowledge Discovery from Sensor Data* (pp. 43–51). ACM.

Rasmussen, J. (1986). *Information processing and human-machine interaction: An approach to cognitive engineering*. New York: Elsevier.

Rubin, D. C. & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*(4), 734–760.

Seibel, R. (1963). Discrimination reaction time for a 1,023-alternative task. *Journal of Experimental Psychology, 66*(3), 215–226.

Sinatra, A. M. & Sottilare, R. A. (2016). Exploring the diversity of domain modeling for training and educational applications. In R. Sottilare, A. Grasser, X. Hu, A. Olney, B. Nye & A. Sinatra (Eds.), *Design recommendations for intelligent tutoring* (pp. 161–164). Orlando, FL: US Army Research Laboratory.

Sottilare, R. A. (2015). Augmented cognition on the run: Considerations for the design and authoring of mobile tutoring systems. In *Foundations of Augmented Cognition* (pp. 683–689): Springer.

Sottilare, R. A. & LaViola, J. (2015). Extending intelligent tutoring beyond the desktop to the psychomotor domain. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2015*. Orlando, FL.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257-285.

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology, 47*, 513–539.

Wickelgren, W. A. (1981). Human learning and memory. *Annual Review of Psychology, 32*, 21–52.

Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin, 76*(2), 92–104.

Wu, P., Zhu, J. & Zhang, J. Y. (2013). Mobisens: A versatile mobile sensing platform for real-world applications. *Mobile Networks and Applications, 18*(1), 60–80.

Zhang, Z., Wu, H., Wang, W. & Wang, B. (2010). A smartphone based respiratory biofeedback system. In *2010 3rd International Conference on Biomedical Engineering and Informatics* (pp. 717–720). IEEE.

CHAPTER 29 – Motivating Individual Difference in an Intelligent Tutoring System

Lauren Reinerman-Jones[1], Elizabeth Lameier,[2] Elizabeth Biddle[3], and Michael Boyce[4]
University of Central Florida[1], Boeing Company[2], US Army Research Laboratory[3]

Introduction

A key ingredient of achievement, engagement, and learning is motivation. Motivated trainees believe, value, focus on learning, manage a task or time more efficiently, and persist. The flip-side results in disengagement, procrastination, anxiety, loss of control, negative thoughts, failure, or a complete shutdown. The benefits to properly assessing and individually motivating learners are saving lives, reducing cost, saving time, and increasing retention. The challenge is determining a goodness of fit for each individual that is measured by an increase in effort, attention, goal attainment, learning outcomes, and retention within an intelligent tutor. Traditionally, trainers develop relationships that paint a clear picture of the individual's motivation, but class size restricts individualization. Intelligent tutors have the potential to assess in real time, plan, and implement individualized motivational strategies to a task lacking luster.

Humans are hunters for the sensation of satisfaction. Boost motivation with targeted reinforcers, aimed at strengthening and magnifying the frequency of a desired response (operant conditioning; Skinner, 1938). Neuroscience notes that the sensation of satisfaction is rewarded with dopamine (DA) being released in the midbrain (ventral stratum). This perhaps begins reinforcement learning (Daw & Shohamy, 2008). According to neuroscience literature, reinforcers increase attention, response rate, ignoring distractors, speed of visual performance, and retention (Small et. al, 2005; Engleman et. al, 2005; Della Libera & Chelazzi, 2009; Wolosin et al., 2012; Murayama & Kuhbander, 2011). Individualized motivation is the golden ticket sought by the Institute for Simulation and Training at the University of Central Florida (IST UCF) in collaboration with Boeing on an effort sponsored by the US Army Research Laboratory, Human Research and Engineering Directorate, Advanced Training and Simulation and Training Division (ARL-HRED-ATSD). One goal for this effort is to plan, develop, and implement a Motivator Assessment Tool to assess motivation for adult learners, matching reinforcers to individuals for increased rate of learning and retention. The Motivator Assessment Tool and reinforcers will then be integrated into the Generalized Intelligent Framework for Tutoring (GIFT; Sottilare, Brawner, Goldberg & Holden, 2012; Sottilare, Brawner, Sinatra & Johnston, 2017).

Past Research

Motivator Assessment Tool Planning

The first step toward individualized motivation is planning a Motivator Assessment Tool. A valuable source of reinforcement learning emerges from special education. Applied Behavior Analysis (ABA) and Functional Behavioral Assessments (FBAs) are the recommended intervention to individualize reinforcers that motivate and change behavior (Blood & Neel, 2007). A measureable goal is created and the problem with ineffective motivation is identified. Often the problem with proper motivation that the student is either avoiding or desiring attention or task completion (Taylor & Abernathy, 2016). Positive reinforcement adds something desired, whereas negative reinforcement (e.g., bad grade) removes something avoided to increase a response. Reinforcers are currently determined for a learner by asking or observing the person (i.e., Premack principle, "If you do this, then this will happen") or using a reinforcer or a preference inventory (e.g., Dunn-Rankin Reward Preference Inventory). Monitoring the effectiveness of the reinforcer notifies

when a modification is required from a declining trendline. Satiation can be avoided by using a selection of reinforcers, varying the pace, and only providing a reinforcer when the desired outcomes are met.

Given that satiation is possible for a single reinforcer, it is important to have multiple reinforcers on-hand for a single learner. Therefore, taxonomies built from diverse contexts should be considered when assessing for reinforcers effective for a learner. Some individuals are motivated by opportunities to socialize, compete, gain recognition, or collaborate in teams, others are motivated by tangibles or non-tangibles. Tangibles can be held by the individual, like badges, money, or gift cards. Non-tangibles are symbols of values, such as digital points or grades that increases a person's attitude and performance (Bari et. al, 2013). Punishment (e.g., loss of point), when used in small portions, balances the expectation of positive reinforcement. In the brain, reinforcers release DA from unexpected rewards, anticipated rewards, and the expectation and value of the reinforcer held by the individual. DA decreases when the reinforcer is not provided or combined with punishment (Daw & Shohamy, 2008).The use of reinforcers provides change in a behavior or motivates the learner, but is hinged on other factors that influence a learner's motivation (Table 1).

Table 1. Factors that influence learner motivation.

Definition	Importance	Assessment Example
Intrinsic motivation: a person's drive to learn is internal and satisfied by knowing, accomplishing, and being stimulated. Extrinsic: applied outside intrinsic motivation.	Intrinsic grows with autonomy, challenge, and competency. Extrinsic motivation for some individuals is regarded as controlling and can be expressed through acceptance, impassiveness, resistance, resentment, and possibly reduces a person's natural intrinsic motivation (Ryan & Deci, 2000). Not every task is intrinsically motivating to the individual.	Intrinsic Motivation Inventory (IMI; Deci & Ryan, 2007), Self-report Scale of Intrinsic versus extrinsic motivation (Harter, 1981), The Academic Motivational Scale (Vallerand et al., 1992)
Values are belief, Schwartz value (1992): power, tradition, conformity, achievement, hedonism, stimulation, self-direction, universalism, benevolence, and security.	The reinforcer needs to hold value to the person to be effective.	The Short Schwartz's value Survey (SSVS; Lindeman & Verkasalo, 2005)
Self-efficacy (Bandura,1997) is a persons' belief about their capability to do a task.	It is rooted in their past success and failures. A person who believes they are capable will maintain effort longer than a learner that does not believe in themselves based on the task.	Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich et al., 1993)
Grit is a person's ability to persist at lifetime or long-term goal despite adversities (Duckworth & Quinn, 2009).	Grit is viewed as an intrinsic motivational quality and therefore, less maintenance of a person is required to keep them motivated.	Brief Grit Scale (Duckworth & Quinn, 2009)
Reinforcement Sensitivity Theory (RST) proposes motivation is changed by a person's sensitivity to rewards and punishments (Gray & McNaughton, 2000).	Knowing a person's sensitivities to reward or punishments will determine strategies used in the plan. For example, punishment may hinder motivation if a person is sensitive to punishment.	Gray-Wilson Personality Questionnaire (GWPQ; Wilson, Gray & Barnett, 1990), BIS/BAS Scales (Carver & White, 1994)
Personality: The Big Five (Goldberg, 1990; Costa & McCrae, 1992) has five traits: extraversion, neuroticism, agreeableness, openness, and conscientiousness.	Different levels of motivation are needed based on the traits of an individuals' personality.	Ten Item Personality Measure (TIPI; Gosling, Rentrow & Swann Jr., 2003)

In addition to those factors described in Table 1, content, procrastination, need for cognition, competence, reward, demographic and interest questionnaires are considered as potential factors in developing the Motivation Assessment Tool. The assessments described previously can be used to obtain an initial baseline of motivation type (intrinsic vs. extrinsic) and determine the types of reinforcers that the learner values. The results determine an initial selection of when and how to include reinforcers to support and encourage student motivation. Inclusion of multidimensional pieces of motivation fortifies the profile given to the intelligent tutor. The development of a new Motivational Assessment Tool to encompass various pieces aimed at strengthening the relationship with the learner and their requirements for learning is the next step.

Proposed Implementation/Enhancement into GIFT

Motivator Assessment Tool Development

Development of the Motivator Assessment Tool began with listing the various reinforcers, taxonomies, and current assessments. This meaningful rudimentary step is the foundation for developing the Motivator Assessment Tool. A document was created that listed questions or statements from current motivational assessments and that content was color-coded. Non-applicable questions or statements for an intelligent tutor were sorted into a discard section of the document. The remaining items were grouped by similarities. For example, related clusters were avoidance, effort, focus, and interest. Clustering the assessments by interrelatedness, guided the streamlining of the new assessment by focusing on multifaceted pieces of motivation.

The next step in developing the Motivator Assessment Tool was to incorporate the relevant factors that were touched on in the previous section. Research provides foundational links for personality with various factors of motivation. A framework was created based on the Big Five personality and includes various associations from grit, values, and RST. For example, a conscientious learner is committed, detailed, organized, linked to high academic success, and exerts effort on both task completion and performance. Grit is associated with conscientiousness because they are dutiful, self-discipline, and achievement striving. (Duckworth,2007). For values, conscientiousness correlates with achievement and security (Parks-Leduc, Feldman & Bardi, 2014). RST positively correlates conscientiousness with sensitivity to being punished (Mitchell & Kimbrel, 2007), and not sensitivity to reward (Mitchell & Kimbrel, 2007; Smits & Broeck, 2006). These constructs and others can be further digested into components and attributes, which can then be connected to motivation and personality. Evaluations of this nature were completed to create a network of nodes captured. Figure 1 illustrates the framework that the intelligent tutor is provided to follow the learner based on personality.

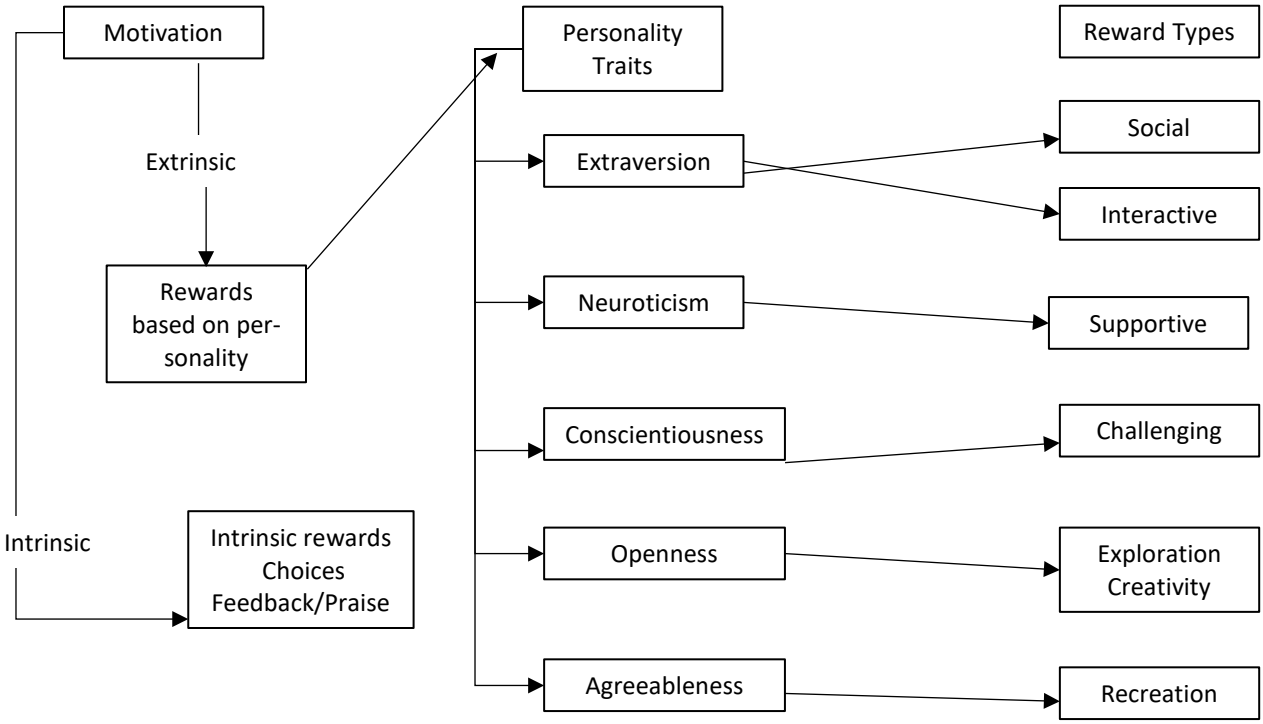


Figure 1. ITS framework based on a learner's personality.

Node Generation Example

For example, conscientious learners tend to be intrinsically motivated, focused, less impulsive, and persist despite obstacles. Conscientious learners care about achievements, compare themselves to others, value getting ahead in life, doing better than others, and being successful. Punishment should be avoided because of their sensitivity, and therefore, the negative effects to motivation. Instead, constructive, positive feedback is required to maintain motivation. The tutor must identify task interest in the learner to avoid suppressing intrinsic motivation. Allowing the learner to choose a more challenging goal and achieving a higher level of achievements suited toward their needs and competence. Rewards are provided, but less frequently because of their own internal drive and detailed nature. If intrinsically motivated, rewards are unexpected and provided at the end of the instructional session. From their detailed perfectionism nature, their stress level might be high and can monitored with physiological measures. The plan for the conscientious learner is significantly altered from a low conscientious learner. On the opposite spectrum, low conscientious learner's characteristics are tied to procrastination, impulsiveness, careless, impatient, and distractible. An association from clustering the assessments finds connections to low conscientious. For example, they are becoming bored easier, give up easier, prone to failure, disposed to low competence/self-efficacy because of their impatient, careless, distractible nature, and low motivation – without adequate support from the ITS. Setting the goal, provide more support to maintain focus, engagement, and motivation for this type of learner. For example, when low motivation is detected, provide a brain break and play motivational music. Supportive continuous guidance through feedback is given to avoid them giving up. A higher level of stimulation is necessary for engagement through pictures, videos, and visuals. This person may need chunked passages and a visual break to help with the amount of time spent focusing on one thing. More praise, more objectives to receive more reinforcers when working toward the goal. The plan to motivate this type of personality is a complete 180 from a person that is conscientious.

Creating the final taxonomy with reinforcers on individual differences is tied to the data collected from the Motivator Tool Assessment and individual differences. Reinforcer selection from the taxonomy finds linking through introverted, extroverted, openness, and whether they are intrinsically motivated. Extrinsic personalities may link to social rewards, competitions, ranking, peer/authority approval, recognition, and helping others. Introverted personalities are associated with independent downtimes, quiet recognition, and spending time with familiar people. Tangible reinforcers are the most difficult items to place because of their dependency on the context, interest, and availability. It is also an item that is used sparingly. Trainers will be responsible for selecting the reinforcers available and execution of the tangible reinforcers.

To test the connections to personality and the new Motivational Assessment Tool, it will be distributed to UCF's student population. The two assessments paired together will create links to individual differences to further validation on previous research and provide new linkage. The Motivator Assessment Tool is the foundation of the project. However, if strong relations are found between personality factors and specific classes of motivators, an additional motivator assessment will not be necessary.

Implementation

Based on phase one, verification of the Motivator taxonomy and/or the Motivator assessment for application with a task in a learning/training environment is addressed in phase two. Presenting the learning objective for application identified jointly by stakeholders in the GIFT platform. This will test how personality and the Motivator taxonomy/assessment, affects the learning rate and retention of a learning objective. For example, if extraversion values the motivational tool of acknowledgement versus introverted preference of free time and relaxation. These factors would then be tied to a more sustained effort, as indicated by physiological measures, such as a higher amount of oxygen produced for a longer sustained time. The goal for phase two is to identify the relations between classifications of motivational tools, different motivational strategies, and individual factors with the learning rate and retention, specifically the Long-Term Learner Model. There will be five scenarios implemented to gather effectiveness on the learner.

Scenario One: Goals and Feedback

Goals and feedback are effective reinforcers and do not interfere with intrinsic motivation (Renninger, 2009). This will begin our baseline of motivation for scenario one. Implementation of goals and feedback will be included across all scenarios due to the recognized effectiveness. In order for feedback to remain effective, it needs be specific, for example, "You crushed moving the basket in a short period of time, awesome accuracy and speed!" Feedback such as knowledge of correct response and elaborated feedback (Timmers, Braber-van den Broek & van den Berg, 2013) have improved learning outcomes.

Scenario Two: Reinforcers Based on Personality

Building off the baseline, the next scenario is the evaluation of personalized reinforcers. In this scenario, individuals would receive the correct reinforcer and others would receive a reinforcer that is yoked with another. This yoking validates if personalization is more effective than any reinforcer provided to the learner. Due to the extensive amount of reinforcers available, validation is made for a select few. Providing insight to effectiveness of individualizing reinforcers based off personality and the effect on learners motivational level.

Scenario Three: Token, Progress, Achievement

As often seen in games, another strategy to reinforce motivation that is a token/achievement economy. The value of the token/achievement economy's efficiency is relatively unknown. This scenario will include a

progress bar, achievement badges with praise, and a token point system. Tokens or points based of achievements a recognizing progress toward meeting the goals and effort. Achievements will also be tied to praise and positive feedback or punishment based on requirements of the personality.

Scenario Four: Intrinsic Motivation

The last section takes in account of intrinsic choices. Participants will have a choice on the level of goal, feedback, and choices such as self-regulation strategies (complexity/visual/chunking). It will provide choice of the level of goal. The levels will be expected goal, above goal, and expert goal. Intrinsic motivation plan with self-regulation tools. Users will be able to adjust the complexity of the text through a tier system that is used in differentiated instruction in the classroom. There will be three levels (high, medium, and low) provided to the learner all lead the learner to accomplishing the expected goal. Complexity could be the vocabulary, amount, type, and examples provided to the learner. The learner can also regulate three levels of chunking on the page to help maintain focus and stress levels. The aim of this scenario is to see if intrinsic learning free from token economies or rewards is superior to retention and the learning rate.

Scenario Five: Reinforcers Combined

The last scenario will have all levels of previous scenarios combined. The first thing the scenario would find is if the learner is intrinsically interested in the task. Initial interest is found by asking a single question about their interest on the topic. The intrinsic learner still may decline in motivation so it would be necessary to ask the question in the middle of the task as well to adjust the plan accordingly based on their personality if a loss of interest has occurred. Even though learners will be provided choices, they will still have the features of a token economy and rewards based on personality. The complete motivational system will be inter connected. For example, if a learner chooses a higher goal the intelligent tutor will also use the token economy to yield a higher return in tokens.

Structuring the scenarios this way, quantify a person's motivation based on the reinforcers used. In theory, the more complete the reinforcer system the more motivational drive a person will have. On top of tailoring motivation specifically to personality and an individual's needs (ex., intrinsic motivation) that enhances the tutor's knowledge of the person and how to motivate them. Different factors may link better with certain types of personalities such as token/ achievement, intrinsic/ self-regulation, and reinforcers and will guide the machine in decisions to manipulate motivation in the learner.

The learning rate will be calculated in the amount of time it took the learner to learn a particular piece of information and the completion of the entire task. Again, in theory, the more motivators provided the quicker the learning will ensue. As well as retention, the more motivation a person has and more release of dopamine, the memory will store the information as an important and place it as schema in the brain. The participants will come back and be assessed on their ability to retain the information from the different sections of reinforcers.

Based on the results of phase 2, the Motivational Assessment Tool with the scenarios presented earlier, another experiment will be run with a different domain population, or scenario for Phase 3. The goal for phase 3 is to identify the validity of the results from the first experiment. Phases 1, 2, and 3 will be used to develop a framework for optimized learning to support the ARL GIFT Long-Term Learner Model Thrust. The results of phases 1, 2, and 3 will form the basis of the framework that will provide pedagogical recommendations based on the evaluation of the student's real-time data on motivation and personality factors into a specific learning intervention in specific GIFT sessions.

Structuring the plan and process differences between different variables with motivation. Being able to see the differences each variable has on an individual versus the whole picture. If every motivational tool is

placed into the system at once, tailoring an individual's needs will never be accomplished. The research may find that certain tools are more effective than others on learning rate and retention. Knowing that each motivational tool effects the learner is important and there hasn't been solid research comparing the different types of motivation on individuals as well as combining and layering factors known to effect motivation to find a good balance for the individual. Research is unclear on the different effects of types of motivation such as tokens/achievements, choices, self-regulation, and reinforcers based on personality. Finding the right amount of certain variables is key to tailoring to an individual. It simply cannot be done without seeking the link of pieces of motivational factors and the whole picture. The intelligent tutor will be able to make decisions on motivators the individual needs if it can predict the effectiveness of that motivational tool for the individual. A perfect fit will be attained with the more knowledge the computer acquires.

Physiological Measures During Implementation

The next step is to monitor and evaluate the effectiveness of the reinforcement strategy on improving and/or maintaining student motivation. This can be accomplished through real-time measures during the instructional session or upon the conclusion of the session. As students are going through the various sessions, they will also be monitored physiologically. Physiological measures can be used to infer affect along two dimensions (Frankenhaeuser, 1986), such as arousal (effort vs. no effort) and stress (euphoria vs. distress), or a discrete emotion, such as fear (Palomba , Sarlo, Angrilli, Mini & Stegagno, 2000). Specific physiological measure used for the effort will be the electroencephalogram (EEG), eye tracking, skin conductance, and electrocardiogram (ECG) to gather baseline data, monitor effect of the reinforcers, and motivational levels throughout the task. Table 2 summarizes the information that can be gleaned from real-time physiological measures during an instructional session.

Table 2. Physiological measure overview.

Physiological Measure	System	Affect Type
Skin Conductance	SNS	Arousal, Engagement, Boredom
Heart Rate	SNS	Arousal, Engagement, Fear
	PNS	Stress, Frustration, Anxiety
EMG	Somatic	Stress, Frustration
EEG	CNS	Workload, Engagement

Regardless of the specific physiological measures used, the change in the physiological measure from baseline to the provision of the motivational reinforcer should be evaluated to make specific interferences regarding the effectiveness of the reinforcer. Additionally, learning performance increase/decrease should also be part of the overall real-time or post-scenario evaluation. Further, physiological responses are also impacted by an individual's personality traits. For instance, introverts tend to become aroused to lower levels of stimuli than extroverts, and individuals high in neuroticism tend to respond with distress to environment events (Eysenck & Eysenck, 1985). The use of these types of measures to make inferences regarding the effectiveness of a reinforcer requires comparing their responses while receiving the reinforcement to a baseline measure taken prior to the start of the instructional session.

Conclusions and Recommendations for Future Research

The next step with motivation on a smaller scale. Reinforcers hold different values to different individuals, while the above project will validate a limited amount of reinforcers, there is a huge list of possible reinforcers needing validation. Assigning a reinforcer an effective value would allow the intelligent tutor to calculate and select the reinforcer to use based on the learners gap in motivation. The gap is the distance

from the actual state to the desired state of motivation. Knowing the level of motivation allows the tutor to make informed decisions on the type of reinforcer needed to optimize the desired state of motivation. It is like a scale, on one end has the actual state and the other end has the desired motivational state. For example, if a person is extremely low on motivation then the machine can say, we need a reinforcer that is going to be powerful and hit with a bang, versus a reinforcer that represents a slight change. The scale may provide information on the amount of times an individual falls below the motivational line and develop a more accurate reinforcement schedule for the individual. Having real-time assessment allows the tutor to make decisions on the type of reinforcer needed to bring them closer to the desired motivational status.

Currently, the intelligent tutoring system is set up with the learners being receivers of information based on their requirements or individual needs. This is a huge advancement to tailoring instruction to an individual; however, discussion with others through teaming and collaboration is also an invaluable source of learning and motivation. The next step for motivation is to allow individuals to discuss and hold powerful conversations to expand their mind, see different perspectives, build prior background knowledge, and experiences. A team established based off of complementary personalities for an effective motivated team should be the next steps in research for implementing into an intelligent tutor.

Everyone is going to come to the table with varying levels of motivation based on the individual. To do nothing keeps a person in the actual state or in a declining state for motivation. Implementing motivational tools, pushes the student toward the desired outcome for motivation. The negative effects to do nothing are astronomical because of the detrimental effects of the emotional, attentional, achievement level it provides. Satisfying a trainee's needs of satisfaction and choice drives them to become intrinsically motivated learners driven by enjoyment of learning. Maintenance of a person who has been motivated is less than a person who is bored, has failed, and has a low self-esteem due to negative thoughts when not motivated.

References

Bandura, A. (1997). *Self-efficacy: the exercise of control.* New York: Freeman.

Bari, N., Arif, U. & Shoaib, A. (2013). Impact of Non-Financial Rewards on Employee Attitude & Performance in the workplace: A case study of Business Institutes of Karachi. *International Journal of Scientific & Engineering Research, 4*(7), 2554–2599.

Blood, E. & Neel, R. S. (2007). From FBA to Implementation: A look at what is actually being delivered. *Education & Treatment of Children, 30*, 67–80.

Cacioppo , J. T. & Petty, R. E. (1982). The Need for Cognition. *Journal of Personality and Social Psychology, 42*, 116–131.

Carver, C. S. & White, T. L. (1994). Behavioral inhibition, behavioral activation and affective responses to impending reward and punishment: the BIS/BAS scales. *Journal of personality and social psychology, 67*(2), 319.

Costa, P. T. & McCrae, R. R. (1992). *NEO-PI-R professional manual.* Odessa, FL: Psychological Assessment Resources

Daw, N. D. & Shohamy, D. (2008). The Cognitive Neuroscience of Motivation and Learning. *Social Cognition, 26*(5), 593–260.

Deci, E.L. & Ryan, R.M. (2007). SDT: Questionnaires: Intrinsic motivation inventory (IMI).

Della Libera, C. & Chelazzi, L. (2009). Learning to attend and to ignore is a matter of gains and losses. *Psychol. Sci.*, 778–784.

Duckworth, A. L., Peterson, C., Matthews, M. D. & Kelly , D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology, 92*(6), 1087.

Duckwork, A. L. & Quinn, P. D. (2009). Development and validation of the Short Grit Scale. *Journal of Personality and Social Psychology, 92*, 1087–1101.

Englemann, J. B., Damaraju, E. & Padmala, S. (2009). Combined effects of attention and motivation on visual task performance: transient and sustained motivational effects. *Front. Hum. Neurosci.*, 3-4.

Eysenck, H. & Eysenck, M. (1985). *Personality and Individual Differences a Natural Science Approach.* New York: Plenum Press.

Frankenhaeuser, M. (1986). A psychobiological framework for research on human stress. In M. Appley & R. Trumbull, *Dynamics of Stress: Physiological, Psychological, and Social Perspectives* (pp. 101–116). New York: Plenum Press.

Goldberg, L. R. (1990). An alternative "description of personality": The Big Five factor structure. *Journal of Personality and Social Psychology, 59*, 1216–1229.

Gosling, S. D., Rentfrow, P. J. & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504–528.

Gray, J. A. & McNaughton, N. J. (2000). *The Neuropsychology of Anxiety 2nd edition.* Oxford: Oxford Medical Publications.

Harter, S. (1982). The Perceived Competence Scale for Childer. *Child Development, 53*(1), 87–97.

Lindeman, M. & Verkasalo, M. (2005). Measuring values with the short Schwartz's value survey. *Journal of personality assessment, 85*(2), 170–178.

Mitchell, J. T., Kimbrel, N. A., Hundt, N. E., Cobb, A. R., Nelson-Gray, R. O. & Lootens, C. M. (2007). An analysis of reinforcement sensitivity theory and the five-factor model. *European Journal of Personality, 21*(7), 869-887.

Murayama, K. & Kuhbandener, C. (2011). Money enhances memory consolidation- but only for boring material. *Cognition, 119*, 120–124.

Palomba , D., Sarlo, M., Angrilli, A., Mini, A. & Stegagno, L. (2000). Cardiac response associated with affective processing of unpleasant film stimuli. *International Journal of Psychophysiology, 36*, 45–57.

Parks-Leduc, L., Feldman, G. & Bardi, A. (2014). Personality traits and personal values a meta-Analysis. *Personality and Social Psychology Review, 19*(1), 3–29.

Pintrich, P. R., Smith, D. A., Garcia, T. & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and psychological measurement, 53*(3), 801–812.

Renninger, K. A. (2009). Interest and identity development in instruction: An inductive model. *Educational Psychologist, 44*(2), 105–188.

Reynolds, W. M. (1988). Measurement of Academic Self-Concept in College Students. *Journal of Personality Assessment, 52*(2).

Ryan, R. M. & Deci , E. L. (2000). Intrensic and Extrensic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, 54–67.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theory and empirical test in 20 countries. *Advances in experimental social psychology, 25*, 1–65.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.* New York: Appleton-Century-Crofts.

Small, D. M., Gitelman, D. & Simmons, K. (2005). Monetary incentives enchances processing in brian regions mediating top-down control of attention. *Cerb. Cortex, 15*, 1855–1865.

Smits, D. J. & Boeck, P. D. (2006). From BIS/BAS to the big five. *European Journal of Personality, 20*, 255-270.

Sottilare, R. A., Brawner, K. W., Goldberg, B. S. & Holden, H. K. (2012). The generalized intelligent framework for tutoring (GIFT).

Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017. DOI: 10.13140/RG.2.2.12941.54244.

Taylor, S. S. & Abernathy, T. V. (2016). Behavior Intervention Flow Chart: A Strategic Tool for Managing Challenging Behaviors. *Scientific Research Publishing, 7*, 2423–2432.

Timmers, C.F., Braber-Van Den Broek, J. & Van Den Berg, S.M. (2013). Motivational beliefs, student effort, and feedback behavior in computer based formative assessment. *Computers & Education,* 60 (1), 25–31.

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C. & Vallieres, E. F. (1992). The Academic Motivation Scale: A measure of intrinsic, extrinsic, and amotivation in education. *Education and Psychological Measurement, 52*(4), 1003–1017.

Wilson, G. D., Gray, J. A. & Barrett, P. T. (1990). A factor analysis of the Gray-Wilson personality questionnaire. *Personality and Individual DIfferences*(11), 1037–1045.

Wolosin, S. M., Zeithamova, D. & Preston, A. R. (2012). Reward modulation of hippocampal subfield activation during successful associative endcoding and retrieval. In: Influence of reward motivation on human declarative memory (Miendlarzewska, Bavelier, Schwartz). *J. Cog. Neuroscience, 24*, 1532–1547.

BIOGRAPHIES

Editors

Dr. Gregory A. Goodwin is a Senior Research Scientist and Acting Branch Chief at the Army Research Laboratory – Human Research and Engineering Directorate at the Paul Ray Smith Simulation and Training Technology Center in Orlando, FL. His research focuses on methods and tools to maximize the effectiveness of training technologies. He holds a Ph.D. in Psychology from Binghamton University and an M.A. in Psychology from Wake Forest University.

Dr. Arthur Graesser is a professor in the Department of Psychology and the Institute of Intelligent Systems (IIS) at the University of Memphis (UofM), as well as an Honorary Research Fellow at University of Oxford. He received his PhD in psychology from the University of California at San Diego. His primary research interests are in cognitive science, discourse processing, and the learning sciences. More specific interests include knowledge representation, question asking and answering, tutoring, text comprehension, inference generation, conversation, reading, education, memory, emotions, artificial intelligence (AI), computational linguistics, and human-computer interaction (HCI). He served as editor of *Discourse Processes* (1996–2005) and is the current editor of the *Journal of Educational Psychology* (2009–2014). His service in professional societies includes president of the Empirical Studies of Literature, Art, and Media (1989–1992), the Society for Text and Discourse (2007–2010), the International Society for Artificial Intelligence in Education (2007–2009), and the Federation of Associations in the Behavioral and Brain Sciences Foundation (2012–2013). In addition to publishing over 600 articles in journals, books, and conference proceedings, he has written 3 books and co-edited 16 books. He and his colleagues have designed, developed, and tested software in learning, language, and discourse technologies, including AutoTutor, AutoTutor-Lite, AutoMentor, ElectronixTutor, MetaTutor, GuruTutor, DeepTutor, HURA Advisor, SEEK Web Tutor, Personal Assistant for Lifelong Learning (PAL3), Operation ARIES!, iSTART, Writing-Pal, Point & Query, Question Understanding Aid (QUAID), QUEST, and Coh-Metrix.

Dr. Xiangen Hu is a professor in the Department of Psychology and Department of Electrical and Computer Engineering at UofM and senior researcher at IIS, and a visiting professor at Central China Normal University (CCNU). He received his MS in applied mathematics from Huazhong University of Science and Technology, MA in social sciences, and PhD in cognitive sciences from the University of California, Irvine. He is the Director of Advanced Distributed Learning (ADL) Center for Intelligent Tutoring Systems (ITSs) Research & Development and a senior researcher in the Chinese Ministry of Education's Key Laboratory of Adolescent Cyberpsychology and Behavior. His primary research areas include mathematical psychology, research design and statistics, and cognitive psychology. More specific research interests include general processing tree (GPT) models, categorical data analysis, knowledge representation, computerized tutoring, and ADL. He receives funding for the above research from the US National Science Foundation (NSF), US Institute for Education Sciences (IES), ADL of the US Department of Defense (DOD), US Army

Medical Research Acquisition Activity, US Army Research Laboratory (ARL), US Office of Naval Research (ONR), UofM, and CCNU.

Dr. Robert A. Sottilare leads adaptive training research at the US Army Research Laboratory where the focus of his research is automated authoring, instructional management, and analysis tools and methods for intelligent tutoring systems (ITSs). He is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT), an open source, AI-based adaptive instructional architecture. He is the lead editor for the Design Recommendations for Intelligent Tutoring Systems book series and the founding chair of the GIFT Users Symposia. He is a program committee member and frequent speaker at the Defense & Homeland Security Simulation, Augmented Cognition, and AI in Education conferences. Dr. Sottilare is a member of the AI in Education Society, the Florida AI Research Society, and the American Education Research Association. He is a faculty scholar and adjunct professor at the University of Central Florida where he teaches a graduate level course in ITS design. Dr. Sottilare is also a frequent lecturer at the United States Military Academy (USMA) where he teaches a senior level colloquium on adaptive training and ITS design. He has a long history of participation in international scientific fora including NATO and the Technical Cooperation Program. Dr. Sottilare is the recipient of the Army Achievement Medal for Civilian Service (2008), and two lifetime achievement awards in Modeling & Simulation: US Army RDECOM (2012) and National Training & Simulation Association (2015).

Authors

Dr. Ryan Baker is an associate professor of education at the University of Pennsylvania. He also has an affiliate appointment at Worcester Polytechnic Institute in the Department of Social Science and Policy Studies and at Teachers College Columbia University, and is also a member of LearnLab and InterLab. He was the founding president of the International Educational Data Mining Society and serves as an associate editor of the *Journal of Educational Data Mining* and the *International Journal of Artificial Intelligence and Education*. His research is at the intersection of educational data mining (EDM) and HCI. He develops and use methods for mining the data that comes out of the interactions between students and educational software to better understand how students respond to educational software and how these responses impact their learning. He studies these issues within intelligent tutors and educational games. In recent years, he and his colleagues have developed automated detectors that make inferences in real time about students' affect and motivational and metacognitive behavior, using data from students' actions within educational software (no sensor, video, or audio data). They have in particular studied gaming the system, off-task behavior, carelessness, "WTF behavior", boredom, frustration, engaged concentration, and appropriate use of help and feedback. They have used these models to make basic discoveries about human learning and learners. Many of these models are developed using data collected through the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) and the HART Android app. He is also the co-developer of the MOOC Replication Framework (MORF).

Dr. Jeffrey M. Beaubien is a principal scientist at Aptima, Inc., where he leads projects on training and human performance assessment in high-risk environments. His research interests include team dynamics, adaptability, and decision making. Dr. Beaubien received a PhD in industrial and organizational psychology from George Mason University, a MA in industrial and organizational psychology from the University of New Haven, and a BA in psychology from the University of Rhode Island.

Dr. Michael W. Boyce is a research psychologist at ARL's Human Research and Engineering Directorate (HRED) supporting the Adaptive Training Research Program under the direction of Dr. Robert Sottilare. His current focus is on investigating the use of adaptive training systems to support the instruction of military tactics and understanding how to properly measure and analyze tactics decision making. His research

interests include leveraging adaptive training technologies to assess tangible user interfaces, as well as understanding how sensors can be used to support tailored learning. He received his doctorate in applied/experimental human factors psychology from the UCL in 2014.

Clayton W. Burford is a science and technology manager within the Advanced Modeling & Simulation Branch of ARL-HRED's Advanced Training Simulation Division (ATSD). He provides technical and programmatic expertise in support of the ARL-HRED-ATSD's advanced distributed simulation research portfolio. He has over 10 years of experience within the Army and Joint community leading technical teams in the fields of M&S and systems engineering.

Zhiqiang Cai is a research assistant professor with the IIS at the UofM. He has a MS degree in mathematics received in 1985 from Huazhong University of Science and Technology, P. R. China. After 15 years of teaching mathematics in colleges, he has worked in the field of natural language processing (NLP) and intelligent systems. He is the chief software designer and developer of Coh-Metrix, OperationAries, CSAL AutoTutor, and many other text analysis tools and conversational tutoring systems. He has co-authored over 70 publications**.**

Dr. Sidney D'Mello is an associate professor with joint appointments in the Departments of Computer Science and Psychology at the University of Notre Dame. His primary research interests are in affective computing, affective science, learning sciences, HCI, and speech and discourse processing. He has co-edited 6 books and has published over 200 journal papers, book chapters, and conference proceedings in these areas. D'Mello and his team have received 11 best/outstanding paper awards at international conferences, have been featured in several media outlets including the *Wall Street Journal*, and have been supported by the NSF; IES; the Gates, Raikes, Templeton, Walton Foundations; Educational Testing Service (ETS), and QUASAR USA. D'Mello served as an associate editor for *IEEE Transactions on Learning Technologies*, *IEEE Access*, and *IEEE Transactions on Affective Computing* (2012–2016). He also served on the editorial boards of the *International Journal of Artificial Intelligence in Education*, *Discourse Processes*, *User-Modeling and User-Adapted Interaction*, *Frontiers in Psychology–Human Media Interaction*, and the *Journal of Educational Psychology* (2011–2013). He serves on the executive boards of the International Artificial Intelligence in Education Society and the Educational Data Mining Society. D'Mello received his PhD in computer science from the UofM in 2009.

Dr. Patrick Donnelly is postdoctoral researcher at the University of Notre Dame. He received a BS and AB from Washington University in St. Louis, an MSE in computer science from Johns Hopkins University, and an MM from the Peabody Conservatory at Johns Hopkins University. Dr. Donnelly received his PhD in computer science from Montana State University in 2015. His primary research interests are focused on EDM and machine learning in the musical domain.

Dr. Peter Foltz is a vice president for research at Pearson and professor adjoint at University of Colorado's Institute of Cognitive Science. Dr. Foltz's research has focused on language comprehension, 21st century skills learning and assessment, and uses of machine learning and NLP in educational technology. The methods he has pioneered improve student achievement, expand student access, and make learning materials more affordable. He has co-lead the framework development for a new assessments of collaborative problem solving and reading literacy for the 21st century for the Organisation of Economic Cooperation and Development's Programme for International Student Assessment (PISA) Assessment. A former professor of psychology at New Mexico State University, he has authored more than 100 journal articles, book chapters, conference papers, and other publications. He previously worked at Bell Communications Research and the Learning Research and Development Center at the University of Pittsburgh. Dr. Foltz holds doctorate and master's degrees in cognitive psychology from the University of Colorado, Boulder, and a bachelor's degree from Lehigh University.

Dr. Jared Freeman is the chief scientist of Aptima, Inc. In that role, he is responsible for aligning Aptima's scientific and technical activities with the company's strategic goals. In his research, Dr. Freeman investigates problem solving by individuals and teams in real-world settings. Dr. Freeman and his colleagues develop decision aids, training systems, measures of performance and communications, and organizational designs that support mission leaders and their staffs. Dr. Freeman is the author of more than 130 articles in journals, proceedings, and books concerning these and related topics. He holds a PhD in human learning and cognition from Columbia University.

Prof. Dr. Samuel Greiff is research group leader, principal investigator, and ATTRACT-fellow at the University of Luxembourg. He holds a PhD in cognitive and experimental psychology from the University of Heidelberg, Germany (passed with distinction). Prof Greiff has been awarded several national and international research funds by diverse funding organizations such as the German Ministry of Education and Research and the European Union (overall funding approx. 9.3 M €), is currently fellow in the Luxembourg research program of excellency, and has published articles in national and international scientific journals and books (>60 contributions in peer-reviewed journals). He has an extensive record of conference contributions and invited talks (>200 talks) and serves as editor for several journals, for instance, as editor-in-chief for the *European Journal of Psychological Assessment*, as associate editor for *Thinking Skills & Creativity*, and as guest editor for the *Journal of Educational Psychology*, *Computers in Human Behavior*, and the *Journal of Business & Psychology*. He has a regular record of ad-hoc reviewing for around 40 different journals and currently serves on five editorial boards. He has been and continues to be involved in the 2012, 2015, and 2018 cycle of PISA, for instance as external advisor to the PISA 2012 and 2015 Expert and Subject-Matter Expert Groups and as contracting partner at his institution. He serves also as chair of the problem-solving expert group for the 2nd cycle of the Programme for the International Assessment of Adult Competencies (PIAAC). In these positions, he has considerably shaped the understanding of transversal skills across several large-scale assessments.

Ross Higashi is a PhD student in learning sciences and policy at the University of Pittsburgh. Previously at Robotics Academy at Carnegie Mellon University (CMU), he worked on the design of CS2N – the computer science student network (http://www.cs2n.org) – and its curricular units. He co-leads the development of the CS2N badge system. He graduated with a BS in logic and computation from CMU, where he also studied computer science and HCI.

Dr. G. Tanner Jackson is a managing research scientist in the Cognitive Science Division at ETS in Princeton, NJ. Dr. Jackson received his PhD in cognitive psychology in 2007 from the UofM and subsequently completed a postdoctoral fellowship there 2008 to 2011. He then continued his work through a position as an assistant research professor within the Learning Sciences Institute (LSI) at Arizona State University (ASU) before joining ETS in 2013. Throughout his career Dr. Jackson has focused his research on adaptive and engaging educational environments that involve elements of game design and NLP components. He has continued that work at ETS where he is involved with developing and evaluating innovative assessments and student process data. His primary ETS projects center around conversation-based formative assessments as well as game-based assessments (working in collaboration with American University's Game Lab). Additionally, he is interested in how users interact with complex systems and leverages these environments to examine and interpret user interactions over time within and across educational systems.

Andy Johnson has been working professionally in distributed learning technology for the last 15 years. He was a developer as the Sharable Content Object Reference Model (SCORM) emerged and was involved in every version of SCORM. He has been working at the ADL Initiative, a research and development unit overseen by the Office of the Under Secretary of Defense for Personnel and Readiness (OUSD(P&R)), for a majority of his career, and is currently serving as the experience API (xAPI) lead. He has designed competency-based content architectures supporting SCORM and xAPI for various government projects, most notably Joint Knowledge Online courses and a series of pharmacy technician training courses designed for

the Services by the Veterans Administration. Mr. Johnson received both his bachelor's degree in computer science and master's degree in education, communication, and technology from the University of Wisconsin-Madison.

Dr. Joan Johnston is a senior scientist with ARL-HRED in Orlando, FL. She is currently leading an Army Science and Technology Objective to develop methods, tools, and strategies for improving the effectiveness of Army simulation training technologies. Dr. Johnston is also the technical lead for a joint project team that is developing an integrated training approach to build resilience, decision making, and teamwork under stress in US Army and US Marine Corps squads. Prior to ARL, she was the Orlando Unit Chief for the Army Research Institute (ARI), leading a team of research psychologists to develop learning principles for employing adaptive training technologies, mobile platform learning environments, and persistent and immersive learning environments. Until 2012, Dr. Johnston was a senior research psychologist and NAVAIR Fellow with the Naval Air Warfare Center Training Systems Division, Orlando, FL. For over two decades, she conducted research on training and decision support systems for tactical decision making under stress, team performance and team training technologies, embedded and distributed simulation-based training technologies, leadership and operational readiness in joint and multinational exercises, and cross-cultural competence. As a principal investigator (PI) and project manager for the ONR-sponsored Tactical Decision Making Under Stress (TADMUS) program, her contributions were rewarded with the ONR Dr. Arthur E. Bisson Prize for Naval Technology Achievement (2000), and the Society for Industrial and Organizational Psychology M. Scott Myers Award for Applied Research in the Workplace (2001). During her career, she has written and presented over 60 professional papers, peer-reviewed journal articles, and presentations.

Dr. Irvin R. Katz is the Senior Research Director of the Cognitive, Accessibility, and Technology Sciences Center (CATS) at ETS. Dr. Katz received his bachelor's degree in computer science from Rensselaer Polytechnic Institute and PhD in cognitive psychology from CMU. Throughout his 27-year career at ETS, he has conducted research at the intersection of cognitive psychology, psychometrics, and technology. His research involves developing methods for applying cognitive theory to the design of assessments, building cognitive models to guide interpretation of test-takers' performance, and investigating the cognitive and psychometric implications of highly interactive digital performance assessments. Dr. Katz is also a HCI practitioner with more than 30 years of experience in designing, building, and evaluating software for research, industry, and Government.

Dr. Sean Kelly (PhD, sociology; University of Wisconsin-Madison) is associate professor and Director of PhD Studies in the Department of Administrative and Policy Studies at the University of Pittsburgh. He studies the social organization of schools, student engagement, and teacher effectiveness. Dr. Kelly's work appears in the *American Educational Research Journal*, *Educational Researcher*, *Teachers College Record*, *Sociology of Education*, *Social Science Research*, and elsewhere. He is the editor of *Assessing Teacher Quality: Understanding Teacher Effects on Instruction and Achievement* (Teachers College Press). In 2014 he received the Exemplary Research in Teaching and Teacher Education award from the American Educational Research Association's (AERA) Division K. He teaches courses in educational reform, leadership, the sociology of education, and statistics for the social and educational sciences. He is currently serving a 2-year position as chair of AERA's Sociology of Education Special Interest Groups. He also serves on the editorial boards of *Educational Evaluation and Policy Analysis*, *Research in the Teaching of English*, *Urban Education*, and the *American Educational Research Journal*.

Dr. Jong Kim is a postdoctoral research fellow at ARL, Orlando, FL. In July 2016, Dr. Kim joined US Army's Intelligent Tutoring Team with the fellowship of Oak Ridge Associated Universities (ORAU). Dr. Kim received his PhD degree in industrial engineering at The Pennsylvania State University, University Park, PA. His research interests lie in the area of cognitive science and engineering. Particularly, Dr. Kim is interested in theories of human learning (forgetting) for the development of intelligent systems. Recently, he developed a theory of skill learning and forgetting (Declarative to Procedural [D2P]) that is being applied

to implement a series of ITSs for the Navy in collaboration with The Pennsylvania State University and Charles River Analytics.

Dr. Bor-Chen Kuo received BS and MS degrees in mathematics education and educational statistics, respectively, from National Taichung Teachers College, Taiwan, R.O.C., in 1993 and 1996, and a PhD degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 2001. He is currently a Distinguished Professor in the Graduate Institute of Educational Information and Measurement, and the Dean of College of Education, National Taichung University of Education, Taiwan. Dr. Kuo is the president of Chinese Association of Psychological Testing and the chief editor of the *Journal of Educational Measurement and Statistics*, Taiwan. He also serves as guest editor in many international journals and an editorial board member of the *Journal of Educational Measurement*. Dr. Kuo received an Outstanding and Excellence Research Award from the R.O.C Education and Research Society in 2009. His research interests include computerized adaptive testing, cognitive diagnostic modeling, machine learning, and AI in education.

Dr. Michelle LaMar is an associate research scientist in the Cognitive, Accessibility, and Technology Sciences Group at ETS. Her current research focuses on the development of psychometric models appropriate for use with complex assessment tasks such as simulations or games. She is particularly interested in modeling task-process data using dynamic cognitive models to enable valid inference about multiple layers of student cognition. Dr. LaMar received a Master's in curriculum studies from Sonoma State University and a PhD in educational measurement from the University of California, Berkeley. Prior to her doctoral work, Dr. LaMar spent 18 years in software engineering, specializing in educational simulations, authoring tools, and natural language parsing.

Elizabeth Lameier, MA, is a research associate in Prodigy at the UCF's Institute for Simulation and Training. Her background stems from the field of education, teaching for almost 10 years in inner city and rural schools. She has a master's in special education and bachelor's in early childhood education.

Dr. Chen-Huei Liao is a professor of special education at National Taichung University of Education in Taiwan. Dr. Liao received her BA from Utah State University, MEd from McGill University, and PhD from University of Alberta from the department of educational psychology in special education. Dr. Liao research interests include the cognitive and non-cognitive factors that facilitate or impede reading acquisition in Chinese, the diagnosis and remediation of Chinese reading difficulties among elementary school children, and the development and application of AutoTutor and Coh-Metrix in Chinese.

Jaclyn Maass is a doctoral student in the psychology department at the UofM. She received her BA in psychology from the University of Tampa and a MSc in general psychology from UofM. Ms. Maass currently works as a research assistant under Dr. Philip I. Pavlik, Jr., in the IIS. Her research is currently focused on creating desirable difficulties during retrieval practice to aid in learning and transfer. Her research interests also include learner modeling, effortful processing, spacing, and individual differences such as motivation and prior knowledge.

Dr. Robert Mislevy is the Frederic M. Lord Chair in Measurement and Statistics at Educational Testing Service and is a professor emeritus at University of Maryland. His 1994 presidential address to the Psychometric Society laid the groundwork for the use of graphical models in educational assessment. He along with co-authors Almond and Steinberg was the recipient of the 2000 National Council on Measurement in Education (NCME) award for Outstanding Technical Contribution to Educational Measurement. Among his honors and awards are the American Educational Research Association's Raymond B. Cattell Early Career Award for Programmatic Research, the National Council of Measurement in Education's Award for Technical Contributions to Educational Measurement (three times), the National Council of Measurement

in Education's award for career contributions to educational measurement, the American Educational Research Association's Lindquist Award for contributions to educational measurement, the ETS Senior Research Scientist Award, and the International Language Testing Association's Samuel J. Messick Memorial Lecture Award. He is a co-author for book *Bayesian Networks in Educational Assessment*, and is co-author with Roy Levy of *Bayesian Psychometric Modeling*.

Dr. Piotr Mitros is the Chief Scientist of edX, an MIT-Harvard educational technology initiative, and is the author of Open edX, an educational platform which has around 300 contributors, and 200 deploys, including ones from the Saudi Ministry of Labor, the Ministries of Education in France, and China, and the Queen Rania Foundation (the not-for-profit of the Queen of Jordan), Stanford, the World Economic Forum, as well as edx.org. As of this writing, it powers around 1,000 full, pure-online courses, has around 10 million users, and forms the backbone of a new research ecosystem. Designed from the ground up for educational data collection, randomized control trials, and experimental pedagogies, at the most recent Learning@Scale conference, all but one of the best paper nominees (including the winner) were based on Open edX. He has been a co-founder or key early employee at three organizations, all of which have crossed the $100 million mark. Dr. Mitros is a frequent conference keynote speaker or panelist on disruption in education, assessment, learning analytics, EDM, open educational resources, and crowdsourcing in education. He has served as an expert on educational policy for the National Academy in Education, the NSF Computing Research Association, and the European Union Commission. Dr. Mitros has taught in China, worked in India facilitated educational technology projects in Nigeria and Jordan, and developed experimental educational formats at MIT. His observations of university systems around the world inspired him to find innovative ways to dramatically increase both the quality of and access to education. He holds a BS in math and electrical engineering, a Master's of Engineering and a PhD in electrical engineering and computer sciences, all from MIT.

Vu Nguyen is a robotics education specialist at the Robotics Academy, a CMU educational outreach program based at the National Robotics Engineering Center in Pittsburgh, PA. He works on the design of CS2N and co-leads the development of its badge system and curricular units.

Dr. Benjamin D. Nye is the Director of Learning Sciences is a research institute at the at the University of Southern California Institute for Creative Technologies (USC-ICT). Dr. Nye's major research interest is to identify best practices in advanced learning technology, particularly for frontiers such as distributed learning technologies (e.g., cloud-based, device-agnostic) and socially situated learning (e.g., face-to-face mobile use). His research interests include modular ITS designs, modeling social learning and memes, cognitive agents, and educational tools for the developing world and low-resource/low-income contexts. He received his PhD in systems engineering from the University of Pennsylvania in 2011. In his recent work as a research professor at the UofM, Dr. Nye led work on the shareable knowledge objects (SKO) framework integrating ITS services such as AutoTutor for the ONR ITS Grand Challenge, helped data mine a corpus of 250k human-to-human online tutoring dialogs (part of the ADL PAL initiative), collaborated on ONR's PAL3 tutoring architecture for supporting life-long learning, and is an advisor and book editor for the ARL GIFT advisory panel. Dr. Nye's research tries to remove barriers development and adoption of ITSs so that they can reach larger numbers of learners, which has traditionally been a major roadblock for these highly effective interventions. He also believes that the future of learning science depends on large, sustainable platforms with many users, where efficient sampling techniques can be used to drive new designs for experiments. Finally, he is interested in making the process of science more efficient, such as by advanced metadata and analysis for scholarly publications.

Dr. Andrew Olney presently serves as associate professor in both the IIS and Department of Psychology and as Director of the IIS at the UofM. Dr. Olney received a BA in linguistics with cognitive science from University College London in 1998, an MS in evolutionary and adaptive systems from the University of Sussex in 2001, and a PhD in computer science from the UofM in 2006. His primary research interests

are in natural language interfaces. Specific interests include vector space models, dialogue systems, unsupervised grammar induction, robotics, and ITSs.

Dr. Kara L. Orvis is the Director of the Performance Assessment and Augmentation Division at Aptima. She is also a principal scientist with over 18 years of experience in government research and development. Her expertise is in the areas of training, leadership, teams, culture, distributed work, and performance measurement for which she has over 70 publications/presentations, including 1 edited book. At Aptima, she leads projects related to military assessment, formation, training, and development. Dr. Orvis holds an MA and PhD in industrial-organizational psychology from George Mason University and a BA in psychology from Ohio Wesleyan University. She is a member of the American Psychological Association and the Society for Industrial and Organizational Psychology.

Dr. Scott Ososky is a postdoctoral research fellow at the Simulation & Training Technology Center (STTC) within ARL-HRED. His current research examines mental models of adaptive tutor authoring, including user experience issues related to development tools and interfaces within the adaptive tutor authoring workflow. He has also published numerous conference papers and book chapters regarding human interaction with intelligent robotic teammates. Dr. Ososky received his PhD and MS in modeling and simulation, as well as a BS in management information systems, from the UCF.

Dr. Philip I. Pavlik is an assistant professor and Director of the Optimal Learning Lab. One mission of the lab is to describe models of learning so that these models can be used by instructional software to sequence and schedule practice. Dr. Pavlik completed his dissertation research with John Anderson in CMU's Psychology Department and has worked with Ken Koedinger in CMU's Human-Computer Interaction Institute. He is current working on multiple existing grants and has applied for funding from both the Department of Education and NSF.

Dr. Lauren Reinerman-Jones is the Director of Prodigy at the UCF's Institute for Simulation and Training. Her lab focuses on assessment for understanding, improving, and predicting human performance and systems. She has over a hundred publications of interdisciplinary work and serves on a variety of Scientific Advisory Boards.

Dr. Vasile Rus is a professor in the Department of Computer Science at UofM with a joint appointment in the IIS. Dr. Rus is also a systems testing research fellow of the FedEx Institute of Technology, a honor received for his pioneering work in the area of software systems testing. His research interests lie at the intersection of AI, machine learning, and computational linguistics with an emphasis on developing interactive intelligent systems based on strong theoretical findings to solve critical challenges that would change the educational and HCI landscape. Dr. Rus has been involved in research and development projects in the areas of computational linguistics and information retrieval for more than 15 years and in open-ended student answer assessment and ITSs for more than 10 years. He has been involved in the development of the following ITSs: DeepTutor (PI), Writing Pal (co-PI), MetaTutor (co-PI), and AutoMentor (co-PI). Dr. Rus has served in various roles on research projects funded by NSF, DOD, and Department of Education, and private companies; has won the first two Question Answering competition organized by the National Institute for Science and Technology (NIST); recently his team won the English Semantic Similarity challenge organized by the leading forum on semantic evaluations – SemEval; has received 4 Best Paper Awards; produced more than 100 peer-reviewed publications; and currently serves as an associate editor of the *International Journal on Tools with Artificial Intelligence* and Program Committee member of the International Conference on Artificial Intelligence in Education (AIED 2015). He is member of the PI Millionaire club at UofM for his successful efforts to attract multi-million funds from federal agencies as PI.

Christian Schunn is a senior scientist at the Learning Research and Development Center and a Professor of Psychology, Learning Sciences and Policy, and Intelligent Systems at the University of Pittsburgh. He

directs a number of research projects in science, mathematics, and engineering education. This work includes studying expert engineering and science teams, building innovative technology-supported science, technology, engineering, and math (STEM) curricula, and studying cognitive and affective factors that influence student and teacher learning.

Dr. David Williamson Shaffer is the Vilas Distinguished Professor of Learning Sciences at the University of Wisconsin-Madison in the Department of Educational Psychology and a game scientist at the Wisconsin Center for Education Research. Before coming to the University of Wisconsin, Dr. Shaffer taught grades 4–12 in the United States and abroad, including 2 years working with the Asian Development Bank and US Peace Corps in Nepal. His MS and PhD are from the Media Laboratory at MIT, and he taught in the Technology and Education Program at the Harvard Graduate School of Education. Dr. Shaffer was a 2008–2009 European Union Marie Curie Fellow. He studies how new technologies change the way people think and learn, and his most recent book is *How Computer Games Help Children Learn*.

Dr. Anne M. Sinatra is an adaptive training scientist at ARL-HRED-ATSD. She works on the GIFT project and is the lead for the Team Modeling for Adaptive Training and Education research vector. Her research interests are focused on cognitive and human factors psychology. She has specific interest in how information relating to the self and about those that one is familiar with can aid in memory, recall, and tutoring. Her dissertation research evaluated the impact of using degraded speech and a familiar story on attention/recall in a dichotic listening task. Her work has been published in the Journal of Interaction Studies, and in proceedings including the Human Computer Interaction International (HCII) Conference and Human Factors and Ergonomics Society (HFES) Conference. She has a combination of over 30 publications and conference papers. Prior to becoming an ARL scientist, Dr. Sinatra was an ARL post-doctoral fellow and graduate research associate with UCF's Applied Cognition and Technology (ACAT) Lab, and taught a variety of undergraduate psychology courses. Dr. Sinatra received her PhD and MA in applied experimental and human factors psychology, as well as her BS in psychology from the UCF.

Dr. Eric Snow is an education researcher in the Center for Technology in Learning at SRI Education. Dr. Snow's current research focuses on the design, development and validation of assessments, particularly performance-based measures of hard-to-assess computational thinking constructs. He has conducted validation studies and developed validation frameworks for performance-based measures of computational thinking, information and communication technology literacy, and teacher candidate's readiness for classroom practice. He has also both led and supported evaluation projects focused on the integration of technology into K–12 contexts, informal science education in after-school contexts and teacher reform. Dr. Snow is currently co-leading a suite of computer science education studies focusing on assessments of learning for secondary school students and on the implementation of a new secondary computer science curriculum as it scales throughout the United States. Dr. Snow earned his PhD in research and evaluation methodology – measurement from the University of Colorado, Boulder, and his BA in anthropology (education studies) from the University of Oregon.

Dr. Erica Snow is a Learning Analytics lead scientist in the Center for Technology in Learning at SRI International. Dr. Snow completed her PhD in cognitive psychology at ASU. Her research explores how data from adaptive technologies can be leveraged to better understand students' cognitive and learning processes.

Dr. Randall D. Spain is a research psychologist with 10 years of experience conducting behavioral science research in both applied and basic research settings. He currently serves as a research psychologist in RTI International's Education and Workforce Development division where he conducts human factors, training, and human performance research for the DOD, Department of Homeland Security (DHS), and Department of Education. His areas of expertise include: human factors and engineering psychology, training design and evaluation, learning and memory, human-automation interaction, and human performance assessment.

Prior to joining RTI, Dr. Spain served as a research psychologist for the US ARI in Orlando, FL, where he designed and validated emerging mobile, virtual, and game-based training capabilities for the US Army. Prior to that he worked as a graduate assistant for ARL where he conducted human factors research with military planning and decision support systems. He received his doctorate in human factors psychology from Old Dominion University. He has received recognition for his work in human factors psychology from the Human Factors and Ergonomics Society and the National Training and Simulation Association.

Dr. Grace Teo is currently a research associate at the IST at UCF. She completed her PhD in applied experimental and human factors psychology from UCF. Her research include assessing workload to inform human-robot teaming, designing vigilance training, understanding decision making, and investigating the effects of individual differences. Before pursuing her PhD, Dr. Teo worked in the civil service and a military research institute in Singapore. Her work in the civil service involved the application of psychometrics and assessment principles in selection and assessment of personnel, as well as in organizational development. At the research institute, her research had more of a human factors focus and included evaluating the effects of new technology on performance, and understanding decision-making processes related to cyber behaviors. From her work in industrial-organizational psychology and human factors psychology, Dr. Teo has extensive experience with a wide range of assessment methods and instruments that extend to the use of various physiological measures. She has published and presented her research in a number of conferences such as the HFES conference, AHFE conference, and the HCII conference. She also has experience teaching both undergraduate and a graduate courses to students of diverse backgrounds.

Dr. Alina von Davier is the Vice President of ACTNext, the ACT, Inc., Research, Development, and Business Innovation Division, as well as an adjunct professor at Fordham University. She earned her PhD in mathematics from the Otto von Guericke University of Magdeburg, Germany, and her MS in mathematics from the University of Bucharest, Romania. At ACT, Dr. von Davier and her team of experts are responsible for developing prototypes of research-based solutions and creating a research agenda to support the next generation for learning and assessment systems (LASs). She pioneers the development and application of computational psychometrics and conducts research on blending machine learning algorithms with the psychometric theory. Prior to her employment with ACT, Dr. von Davier was a Senior Research Director at ETS, where she led the Computational Psychometrics Research Center. Previously, she led the Center for Psychometrics for International Tests, where she managed a large group of psychometricians, and was responsible for both the psychometrics in support of international tests, TOEFL® and TOEIC®, and the scores reported to millions of test-takers annually. Two of her volumes, a co-edited volume on *Computerized Multistage Testing*, and an edited volume on test equating, *Statistical Models for Test Equating, Scaling, and Linking*, were selected, respectively, as the 2016 and 2013 winners of the AERA Division D Significant Contribution to Educational Measurement and Research Methodology award. In addition, she wrote or co-edited five other books and volumes on statistic and psychometric topics. Her current research interests involve developing and adapting methodologies in support of virtual and collaborative learning and assessment systems. Machine learning and data-mining techniques, Bayesian inference methods, and stochastic processes are the key set of tools employed in her current research. She serves as an associate editor for *Psychometrika* and the *Journal of Educational Measurement*. Prior to joining ETS, she worked in Germany at the Universities of Trier, Magdeburg, Kiel, and Jena, and at the ZUMA in Mannheim, and in Romania, at the Institute of Psychology of the Romanian Academy.

Dr. Duanli Yan is a manager of data analysis and computational research for Automated Scoring group in the Research and Development Division at ETS. She is also an adjunct professor at Rutgers University. She has been working on Bayesian inference networks, Bayesian analysis with Markov Chain Monte Carlo and evidence-centered design in educational assessment. Some of her work can be found in *Bayes Nets in Educational Assessment: Where the numbers come from?* (Mislevy, Almond and Yan, 2000*).* She received many awards including 2011 ETS Presidential Award, 2013 NCME Brenda Loyd award, and 2015 IACAT Early Career Award. She is a co-author for book *Bayesian Networks in Educational Assessment*. She is also

a co-editor for volume *Computerized Multistage Testing: Theory and Applications*. She has been an invited symposium organizer and presenter at many conferences including NCME, International Association for Computerized Adaptive Testing (IACAT), and International Psychometrics Society (IMPS).

Dr. Louise Yarnall, an education researcher, is a senior social science researcher at SRI International, Menlo Park, CA; she specializes in STEM workforce training and assessment design, adult learning and motivation, and designing innovative technologies for lifelong learning. She conducts learning science literature reviews, cognitive task analyses, and performance assessment design and implementation on various DOD education and training initiatives. She holds a PhD in education from the University of California, Los Angeles.

Dr. Diego Zapata-Rivera is a senior research scientist in the CATS Center at ETS in Princeton, NJ. He earned a PhD in computer science (with a focus on AI in education) from the University of Saskatchewan in 2003. His research at ETS has focused on the areas of innovations in score reporting and technology-enhanced assessment including work on adaptive learning and assessment environments, and conversation-based and game-based assessments. His research interests also include Bayesian student modeling, open student models, virtual communities, authoring tools, and program evaluation. Dr. Zapata-Rivera has produced over 100 publications including journal articles, book chapters, and technical papers. He has served as a reviewer for several international conferences and journals. He has been a committee member and organizer of international conferences and workshops in his research areas. He is a member of the editorial board of *User Modeling and User-Adapted Interaction* and an associate editor of the *IEEE Transactions on Learning Technologies Journal*. Most recently, Dr. Zapata-Rivera has been invited to contribute his expertise to projects sponsored by the National Research Council, the NSF, and NASA.

INDEX

adaptive instruction, 17, 41, 320, 326
adaptive training, 10
adaptive tutoring, 5, 10
Albert, i, iv, 18, 21, 22, 23, 24, 26, 27, 28, 260, 309, 315, 317, 327
architecture, 3, 9
assessment, 3, 6, 9, 10, 11, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 33, 36, 39, 41, 43, 44, 45, 47, 48, 49, 50, 53, 54, 55, 56, 57, 59, 61, 64, 65, 66, 69, 71, 73, 75, 76, 77, 80, 81, 82, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 99, 100, 101, 102, 103, 104, 105, 106, 108, 109, 119, 120, 121, 122, 123, 125, 126, 129, 131, 132, 133, 134, 135, 137, 138, 140, 141, 147, 148, 151, 152, 153, 155, 157, 159, 160, 164, 165, 166, 167, 171, 174, 176, 177, 178, 183, 184, 185, 186, 187, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 203, 204, 205, 206, 207, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 229, 230, 231, 232, 236, 237, 238, 239, 240, 242, 243, 244, 246, 247, 249, 250, 251, 252, 253, 254, 255, 256, 259, 260, 265, 268, 270, 271, 272, 273, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 287, 288, 290, 291, 295, 297, 299, 301, 303, 304, 305, 306, 310, 312, 314, 315, 316, 317, 318, 319, 320, 326, 328, 333, 335, 338, 339, 342, 343, 346, 347, 348, 349, 350, 351
authoring tools, 6, 7
Ayers, iv, 260, 288, 292
Baker, ii, 3, 11, 20, 43, 45, 60, 66, 70, 81, 82, 91, 92, 94, 95, 96, 100, 104, 121, 123, 135, 137, 142, 143, 146, 152, 153, 172, 174, 176, 177, 178, 179, 192, 193, 200, 216, 224, 246, 276, 284, 314, 315, 318, 342
Beaubien, iii, 185, 342
Biddle, iv, 260
Bink, iv, 260, 287, 288, 292
Boyce, iv, 260, 342
Brawner, ii, 3, 4, 8, 11, 12, 17, 32, 33, 39, 95, 100, 131, 133, 135, 136, 167, 207, 222, 224, 231, 233, 236, 246, 247, 283, 285, 287, 292, 293, 295, 305, 306, 317, 318, 331, 339
Burford, iii, 185, 343
Cai, iv, 168, 260, 279, 282, 283, 284, 301, 304, 305, 306, 310, 314, 315, 317, 318, 343
D'Mello, iv, 3, 9, 11, 94, 159, 163, 169, 170, 176, 178, 224, 259, 264, 271, 272, 273, 274, 343
Diedrich, iv, 260, 287, 292
domain knowledge, 8
domain model, 3, 5
domain module, 3, 6
Donnelly, iv, 259, 264, 272, 343
evidence model, 73, 77, 100, 104, 105, 126, 133, 134
evidence rules, 104, 126
Evidence-Centered Design (ECD), 73, 102
expert model, 8
expertise, 9, 11
Folsom-Kovarik, i, 18
Foltz, i, ii, iii, iv, 18, 100, 157, 160, 162, 169, 170, 174, 178, 183, 193, 200, 260, 275, 279, 280, 283, 284, 309, 314, 317, 318, 343
Freeman, iii, 27, 185, 226, 233, 338, 344
Gašević, ii, iii, 94, 95, 100, 177, 179, 183, 200, 246
Generalized Intelligent Framework for Tutoring, ii, iii, iv, 4, 12, 17, 29, 39, 44, 50, 69, 92, 99, 101, 125, 131, 137, 167, 172, 178, 183, 201, 207, 210, 225, 233, 236, 247, 249, 256, 259, 274, 276, 288, 292, 293, 295, 310, 326, 331
GIFT, i, ii, iii, iv, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 17, 18, 19, 20, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 44, 50, 69, 80, 92, 95, 99, 100, 101, 119, 125, 131, 132, 133, 134, 136, 137, 151, 167, 172, 177, 178, 183, 184, 185, 186, 187, 200, 201, 207, 210, 222, 225, 231, 233, 236, 239, 240, 241, 242, 244, 245, 246, 247, 249, 250, 251, 252, 253, 254, 255, 256, 259, 260, 270, 271, 274, 276, 279, 282, 283, 288, 290, 292, 293, 295, 297, 301, 303, 304, 305, 306, 310, 316, 326, 331, 333, 335, 336, 339, 347, 349
GIFT Authoring Function, 7
Goldberg, ii, 3, 4, 8, 10, 11, 12, 13, 20, 32, 33, 39, 41, 44, 45, 93, 95, 131, 136, 167, 189, 193, 201, 207, 208, 231, 233, 236, 246, 247, 255, 256, 287, 292, 293, 295, 306, 309, 318, 326, 328, 331, 332, 339
Goodwin, 3, i, iii, iv, 18, 33, 39, 181, 183, 186, 260, 283, 317, 341
Graesser, 3, ii, iv, 3, 9, 11, 12, 17, 20, 69, 82, 89, 95, 97, 99, 127, 135, 136, 157, 159, 161, 170, 171, 173, 176, 178, 179, 193, 200, 216, 223, 224, 260, 273, 274, 275, 276, 279, 280, 282, 283, 284, 285, 292, 295, 301, 304, 305, 306, 309, 310, 311, 314, 315, 317, 318, 341
Greiff, ii, iii, iv, 100, 137, 139, 141, 153, 184, 193, 200, 221, 224, 260, 276, 278, 284, 344
Higashi, i, 19, 53, 66, 344
Hu, 3, i, ii, iii, iv, 11, 18, 19, 27, 41, 42, 45, 69, 82, 95, 100, 136, 159, 170, 171, 176, 178, 179, 189, 201, 224, 257, 259, 260, 279, 281, 283, 284, 285, 292, 295, 297, 299, 300, 301, 302, 304, 305, 306, 309, 310, 311, 314, 315, 317, 318, 329, 341

intelligent tutoring system, 3, 4, 5, 6, 7, 8, 9, 10, 11
Jackson, ii, 100, 127, 135, 136, 216, 222, 223, 224, 275, 285, 309, 310, 317, 318, 344
Johnson, i, 3, 9, 12, 18, 19, 24, 27, 29, 30, 39, 104, 122, 127, 134, 135, 344
Johnston, iii, 3, 5, 13, 17, 39, 185, 345
Katz, ii, iii, 19, 73, 82, 100, 127, 136, 183, 184, 210, 216, 217, 221, 222, 223, 224, 252, 275, 285, 309, 318, 345
Kelly, iv, 259, 261, 263, 272, 273, 274, 338, 345
Kim, iii, iv, 94, 184, 193, 195, 201, 260, 320, 321, 328, 345
Kuo, i, iv, 18, 260, 278, 284, 346
LaMar, ii, iii, 100, 148, 153, 184, 346
Lameier, iv, 260, 331, 346
learner model, 3, 5, 8, 10
learner module, 3, 6
Liao, iv, 260, 284, 346
Maass, iii, 184, 346
Mislevy, ii, 44, 45, 54, 66, 69, 74, 75, 82, 86, 95, 100, 101, 103, 104, 106, 107, 109, 120, 121, 122, 123, 125, 128, 134, 135, 174, 178, 285, 346, 350
Mitros, ii, iii, 20, 85, 90, 91, 95, 178, 183, 193, 201, 347
Nguyen, i, 19, 208, 325, 328, 347
Nussbaumer, i, 18, 25, 26, 27, 28
Nye, i, iii, 19, 41, 42, 45, 171, 178, 183, 189, 194, 197, 201, 261, 273, 284, 295, 301, 304, 305, 306, 311, 315, 318, 329, 347
Olde, i, 19, 47, 50
Olney, ii, iv, 3, 12, 95, 100, 163, 169, 223, 224, 259, 264, 272, 273, 274, 280, 284, 295, 306, 310, 314, 317, 318, 329, 347
Orvis, iii, 185, 187, 348
Ososky, i, iii, 19, 65, 67, 185, 207, 236, 246, 292, 348
pedagogical model, 3, 8
pedagogical module, 3, 6
pedagogy, 7
Reinerman-Jones, iii, iv, 185, 260, 348
Rus, ii, 3, 12, 100, 155, 156, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 279, 280, 281, 284, 285, 314, 318, 348
Schatz, i, 18
Schunn, i, iii, 19, 53, 62, 66, 67, 183, 191, 192, 200, 201, 348
Shaffer, iv, 189, 200, 260, 279, 280, 281, 283, 284, 285, 349
Sinatra, iii, 3, 5, 13, 17, 39, 133, 135, 185, 186, 207, 236, 247, 255, 256, 271, 274, 288, 292, 295, 306, 319, 320, 329, 349
Snow
Eric, ii, 19, 75, 81, 212, 221, 223, 224, 318, 349
Erica, ii, 19, 349
Sottilare, 3, i, ii, iii, iv, 1, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 15, 17, 18, 20, 32, 33, 39, 44, 45, 69, 82, 95, 131, 133, 135, 136, 167, 171, 176, 178, 179, 185, 193, 201, 207, 208, 224, 231, 233, 236, 246, 247, 260, 275, 279, 283, 285, 287, 288, 292, 293, 295, 305, 306, 309, 317, 318, 319, 320, 324, 326, 329, 331, 339, 342
Spain, iii, 27, 135, 178, 184, 207, 349
Teo, iii, 185, 350
tutor model, 3, 4
user interface, 3, 7
von Davier, ii, 100, 122, 136, 174, 178, 217, 218, 223, 224, 350
Yan, ii, 100, 101, 121, 129, 135, 328, 350
Yarnall, ii, 19, 351
Zapata-Rivera, i, ii, iii, 19, 42, 45, 100, 125, 127, 129, 135, 136, 184, 216, 217, 218, 222, 223, 224, 275, 279, 284, 285, 309, 318, 351

Design Recommendations for Intelligent Tutoring Systems

Volume 5
Assessment Methods

Design Recommendations for Intelligent Tutoring Systems (ITSs) explores the impact of intelligent tutoring system design on education and training. Specifically, this volume examines "Domain Modeling". The "Design Recommendations book series examines tools and methods to reduce the time and skill required to develop Intelligent Tutoring Systems with the goal of improving the Generalized Intelligent Framework for Tutoring (GIFT). GIFT is a modular, service-oriented architecture developed to capture simplified authoring techniques, promote reuse and standardization of ITSs along with automated instructional techniques and effectiveness evaluation capabilities for adaptive tutoring tools and methods.

About the Editors:

- ***Dr. Robert Sottilare*** *leads adaptive training research at the Army Research Laboratory and is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT).*
- ***Dr. Arthur Graesser*** *is a professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis and is a Senior Research Fellow in the Department of Education at the University of Oxford.*
- ***Dr. Xiangen Hu*** *is a professor in the Department of Psychology at The University of Memphis and visiting professor at Central China Normal University.*
- ***Dr. Gregory Goodwin*** *is a senior adaptive training scientist at the U.S. Army Research Laboratory.*

A Volume in the Adaptive Tutoring Series

978-0-9977257-2-8