Int J Artif Intell Educ
DOI 10.1007/s40593-017-0155-y

ARTICLE

Modeling Expert Behavior in Support of an Adaptive Psychomotor Training Environment: a Marksmanship Use Case

Benjamin Goldberg[1] · **Charles Amburn**[1] · **Charlie Ragusa**[2] · **Dar-Wei Chen**[3]

Abstract The U.S. Army is interested in extending the application of intelligent tutoring systems (ITS) beyond cognitive problem spaces and into psychomotor skill domains. In this paper, we present a methodology and validation procedure for creating expert model representations in the domain of rifle marksmanship. GIFT (Generalized Intelligent Framework for Tutoring) was used as the architecture to guide development efforts and was paired with an Army marksmanship simulator that collects behavioral information through sensor technologies. The models were based on expert data from eight members of the U.S. Army Marksmanship Unit's Service Rifle Team. The goal is to establish validated models that serve as artificial intelligence assessment criteria for driving a self-regulated training environment. We review the techniques applied to the data for model construction, the trends found in the data that are generalized across each expert informed through cross-fold validation practices, and discuss how the models will be used for driving real-time assessment. Results support the utility of generalized expert models across the fundamental components of rifle marksmanship as outlined in U.S. Army doctrine.

Keywords Expert modeling · Psychomotor · Marksmanship · GIFT · Intelligent tutoring systems

✉ Benjamin Goldberg
benjamin.s.goldberg.civ@mail.mil

Charles Amburn
charles.r.amburn.civ@mail.mil

[1] U.S. Army Research Laboratory, Orlando, FL, USA

[2] Dignitas Technologies, Inc., Orlando, FL, USA

[3] Georgia Institute of Technology, Atlanta, GA, USA

Published online: 12 September 2017

Introduction

The United States Army seeks to leverage advancements in computing technologies to reshape how simulation-based training applications are used. Ultimately, the goal is to develop adaptive digitized learning products that employ artificial intelligence and/or digital tutors to tailor learning to the individual soldier across an array of military relevant domains (Department of the Army 2011), with a goal of replicating findings highlighted in Bloom's 2-sigma problem (Bloom 1984). The aim is to develop technologies that produce the same learning benefits achieved when individuals acquire knowledge and skill through interactions with an expert human tutor or coach. While Intelligent Tutoring Systems (ITS) have seen success among multiple academic and military applications (Kulik and Fletcher 2016; VanLehn 2011), most triumph stories are seen in well-defined cognitive problem spaces (e.g., VanLehn 2011; Stottler et al. 2001; Steenbergen-Hu and Cooper 2014; Steenbergen-Hu and Cooper 2013). A recognized gap in this training paradigm is the soldier's requirement to perform critical tasks in domains that involve interplay between cognitive and psychomotor task components. Prior work has investigated the application of ITSs for training procedural mechanical tasks within virtual reality scenarios guided by pedagogical agents (Rickel and Johnson 1999), but their utility has not extended into the physical task environment. In lieu of this, Lieberman and Breazeal (2007) investigated a vibrotactile feedback approach to learning new motor skills, with improved accuracy when that channel of feedback was active. The findings across this study are promising; however, their approach outside of controlled lab settings is still unproven.

Creating an ITS that accounts for a combination of cognitive and psychomotor elements can prove challenging. This is because a system of this nature must be able to collect and model task relevant behavioral measures for the purpose of guiding formative assessments. In addition, it must do so in the tasks' natural environment without hindering task-related movements of an individual (Goldberg 2016). This ensures models derived from interaction are not impacted by the means in which the behavioral data is acquired. Rather than modeling and monitoring steps toward solving a problem and identifying misconceptions and impasses along the way, the psychomotor use case focuses on behavior and its inherent influence on performance; specifically, what nuances of a task are dictated by behavioral patterns and what strategies can be enacted to assist an individual in acquiring the ability to replicate a desired behavior across multiple trials. Investigating technologies to facilitate this inference procedure is critical. A psychomotor ITS requires methods to consume perceptual information that associates with task relevant behaviors, along with models that determine how the captured data relate to a representation of desired behavior or common errors (i.e., expert models and buggy-libraries). Collection of such data is technologically difficult to capture and classify in a manner that is directed for assessment purposes.

In an effort to evaluate the efficacy of utilizing ITS methods for training psychomotor skills, a military relevant domain was selected to guide development efforts. From a programmatic perspective, we sought a domain currently using sensors to collect behavior and performance data to assist in the development of a set of physical skills. With this criterion, the domain of rifle marksmanship was selected, as it is of the highest priority to the U.S. Army and currently utilizes high-end custom

simulators with embedded sensors in their program of instruction (Ranes et al. 2014). Marksmanship is an excellent initial use case as it is a complex psychomotor skill demanding high levels of concentration and extended applications of fine motor control (Pojman et al. 2009). In addition, it is considered a pivotal fundamental to success of military operations (Yates 2004).

Basics of Rifle Marksmanship

For building an adaptive training capability, it is important to understand the nuances associated with marksmanship training. In terms of objectives, rifle marksmanship focuses on the basic functional elements for effectively operating the weapon, and begins with instruction on consistently striking static targets at fixed distances (Department of the Army 2016). The Engagement Skills Trainer II (EST II) is a simulated firing range developed to provide a cost-saving solution for deliberate practice of marksmanship functional elements (Department of the Army 2016). It was primarily intended to help meet high throughput requirements while controlling cost on ammunition and time associated with live range exercises.

The EST II was selected for our research because of its utility to collect behavioral data at a granular enough level to inform model representations for driving assessment. The simulated weapons utilized in the EST II are manufactured with embedded sensors that captures behavioral data during system interaction (e.g. trigger pull, weapon orientation, and aim trace). In practice, these sensor features were included as information channels an instructor could monitor to better diagnose shooter errors. This comes with the assumption that an instructor can efficiently access this data and make accurate judgments on what an individual is doing incorrectly. Though proper usage has been shown to increase performance while minimizing cost and risk to soldiers (Platte and Powers 2008), James and Dyer (2011) point out that human instructors can be subjected to high workload due to concurrent monitoring of multiple trainees and can be inconsistent in their interpretation of the data and in their capacity to provide effective coaching. Thus, the primary goal of this research is to examine how artificial intelligence (AI) methods can be applied in this context to drive objective assessments that are data-driven and void of subjective inconsistencies related to limitations of human coaches attending to multiple channels.

Building an Adaptive Marksmanship Capability

From an implementation standpoint, applying intelligent tutoring to a psychomotor domain requires the same piece parts associated with developing any ITS. At a minimum, this includes (1) data types at a granular enough level to inform appropriate assessments, (2) established models of expert performance to inform performance state determinations, and (3) a pedagogical model that guides practice and accelerates skill acquisition through formative feedback and adaptive sequencing (Goldberg et al. 2015). With advancements in sensing technologies that capture a variety of behavioral markers, creating a psychomotor ITS is now more achievable than ever (Goldberg 2016).

In the case of building an adaptive tutor for rifle marksmanship in the EST II, the initial task is establishing models that designate performance outcomes across the set of behavioral data sources made available from the simulator (Amburn et al. 2014). In this

instance, we are using the Rifle and Carbine Training Circular, (TC) 3-22.9 (Department of the Army 2016) as guidance to match data sources with specific functions of the shot process as taught by the U.S. Army. This TC is an Army generated document that provides specific information on the carbine and how it functions, its capabilities, and the application of the functional elements of the shot process. The functional elements (stability, aiming, control, and movement) identified in the TC are based on thorough task analyses that define the critical aspects of operating a rifle, with explicit recommendations on how best to perform those functions. In this research, we focus on selected elements of control (i.e., trigger squeeze and breath control) and body stability, based on available data from the simulator and the goal of coaching grouping procedures alone. Historically, these specific elements represented three of the four fundamentals of marksmanship behavior within Army doctrine. Our approach to assessing the fourth element, aiming, will be reviewed in the discussion. While the TC provides a qualitative representation of proper fundamental techniques (top-down), data is required to back-up these assertions with quantitative methods that provide an objective-based approach to behavioral performance assessments (bottom-up). Similar approaches have been applied to expert model development (Ritter and Feurzeig 1988), but few studies have been applied to determine the effectiveness of such approaches.

In relation to this domain, prior work has investigated data-driven assessments in marksmanship training. Using a sensor-embedded weapon simulator like those in the EST II, Chung et al. (2009) were able to improve shooting skills by providing individualized instruction linked to errors in performance (shot placement) and behavior (body position, breathing, trigger squeeze and muzzle wobble). In that study, a human coach was used to review the data and diagnose errors based on a checklist provided by the researchers. While the checklist was provided to assist in diagnosing error, the subjective nature of a human coach introduces the potential for inconsistency. In contrast, an ITS could autonomously review the data, diagnose the problems, and prescribe and deliver relevant feedback based on a unified representation of appropriate behavior. To accomplish this, the ITS requires the capability to analyze the shooter data and determine how the represented behaviors compare to Army standards. Thus, a critical step in developing the logic to drive the adaptive marksmanship training system was to collect performance data from expert marksmen and use that data to create an expert model for use in evaluating shooter performance in real-time. In a complementary study, Nagashima et al. (2008) examined sensor-based measures for determining differences in marksmanship performance and expertise levels. Their resulting regression models found breath and trigger control as significant predictors in classifying an individual as novice or expert. However, their approach does not provide a real-time assessment capability at the functional element level required for ITS application and focused coaching practices.

In the following sections, we describe the process employed to create models of rifle marksmanship based on expert data from members of the U.S. Army Marksmanship Unit's (AMU) Service Rifle Team. For development purposes we leveraged GIFT (Generalized Intelligent Framework for Tutoring), an architecture-based project providing the tools and methods for authoring and delivering adaptive training in a variety of instructional domains (Sottilare et al. 2012). For this project, GIFT has been configured to receive the sensor data coming from the marksmanship simulator, and physiological sensors worn by the shooter. We present the modeling techniques applied, the trends discovered, and the methods used for validation purposes. Further,

we discuss how the models, used by GIFT, can serve as the AI assessment criteria for providing autonomous, adaptive training.

Methodology

Participants

The expert marksmen for this study were recruited from the U.S. AMU Service Rifle Team stationed at Fort Benning, GA. Based on previous expert modeling efforts conducted by a leading researcher in the field at the University of Central Florida, Dr. Avelino Gonzalez, the sample size selected was eight. This number is sufficient for exploratory analyses that determine if a generalized model of expert performance can be constructed. Dr. Gonzalez said, based on his experience and those of colleagues, that once you get past a few experts they all start doing the same thing, the same way. This is supported by a similar field of usability study, where on average five experts identify up to 80% of usability problems (Nielsen 1994). You just have to get a few experts' worth of data if they are truly top performers in the field and then validate the models to see if they hold up. Of the eight experts, six were male and two were female, with an average age of 26. For reference, the average size of the AMU Service Rifle Team is 16, with small fluctuations in composition from year to year. Two members were repeat national champions in service rifle competitions, and five were recognized as being members of the top one-hundred shooter's in the country, having received the President's Hundred Tab (a badge awarded by the Civilian Marksmanship Program to the 100 top-scoring military and civilian shooters in the annual President's Rifle Match).

Apparatus

The experimental setup consisted of hardware and software components associated with Meggitt's Fire Arms Training System (FATS) M100 Advanced Reality Simulator (Meggitt Training Systems n.d.). Similar to the Army's EST II, this apparatus includes a simulated M4 carbine with embedded sensor technologies for monitoring behavior variables. In addition, a physiological sensor was incorporated to collect data associated with breathing patterns not inherently captured by the FATS M100. Each component is described in detail below.

Marksmanship Training Simulator

The FATS M100 is a simulated marksmanship training environment that supports individual and collective training events across a full range of weapons. The system is composed of four components: (1) a computer hard-drive containing all simulation components; (2) a projection system to visualize marksmanship ranges; (3) a hit detect camera used to locate shot placements; and (4) the simulated weapon.

The system operates through an infrared laser mounted directly within the barrel of the rifle. The shooter aims at digital targets projected on a screen. The system

logs the point of aim in real-time as measured through the optical sensor in the hit detection camera tracking the laser. When the weapon's trigger is activated, the ballistic flyout is computed and the point of impact is recorded. The performance measures were computed locally on the FATS M100 and sent to the ITS framework for logging.

The FATS M100 was configured with a virtual shooting range consisting of a single firing lane and a standard 25 m zeroing target (see Fig. 1). The physical distance from the firing line to the screen was 26 ft, with the system performing all the required scaling to simulate the 25 m range. Although the target is placed (physically in the real world, digitally on the simulator) 25 m away from the firing line, the human silhouette on the target is scaled down to represent a target that is 300 m away. Automatic dispersion and wind effects were disabled for the data collection.

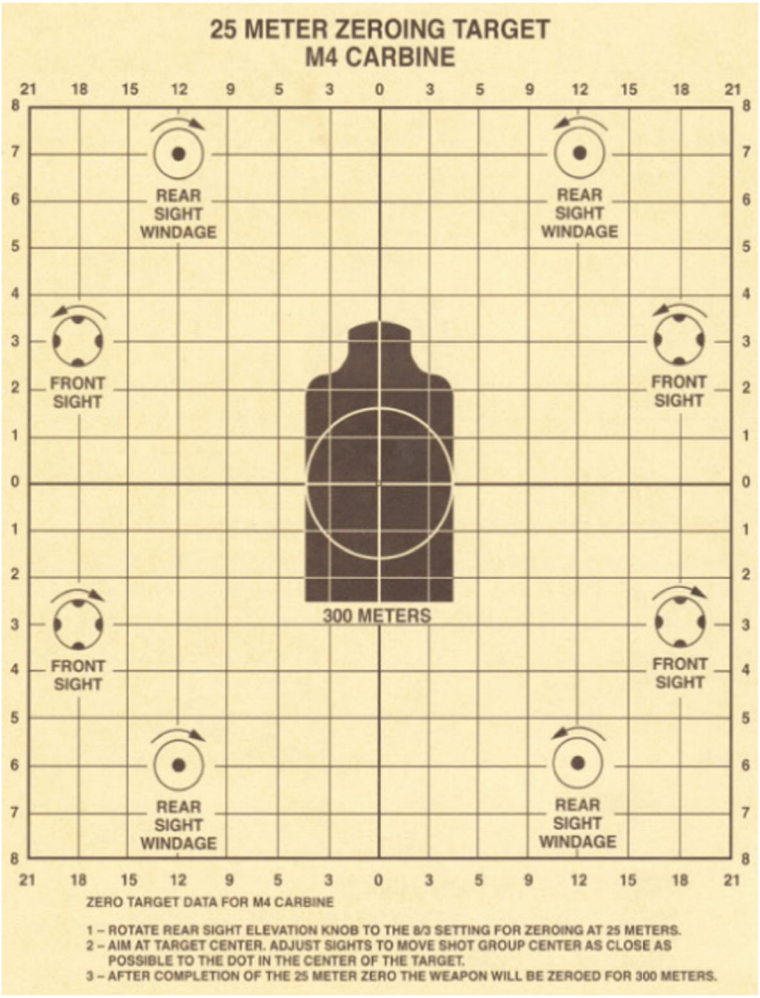

Fig. 1 Standard U.S. Army 25 m zeroing target

Instrumented Weapon

The weapon used for this study is Meggitt's fully-sensored M4 carbine that is laser aligned and assembled specifically for simulator use. It has been validated by the U.S. Army to have the form, fit, and function of an actual standard issue M4. The weapon was designed with embedded sensors, including: point of aim through an infrared laser mounted in the barrel at a 6 Hz sampling rate, a trigger sensor that measures displacement during the execution of a shot, butt-stock pressure, and the weapon's cant angle. All sensors were built within the form of the weapon and were not discernible to users. Each sensor stream was logged separately and used as the primary inputs in model development. For initial modeling efforts, weapon cant and butt-stock pressure were not analyzed, and will not be reported upon in this paper. To keep the effort manageable, the first-pass at the adaptive marksmanship trainer was to focus on models linked to the functional elements as outlined in the Army TC (Department of the Army 2016), with barrel movement, trigger, and breathing serving as the model inputs.

It is worth noting that the carbine was tethered by a long cable plugged directly into the FATS M100 system for data logging, as well as into an air compressor that created recoil effects in the weapon. The amount of recoil is not the intensity of a real weapon fired (approximately 70%), but sufficient to require a trainee to manage the recoil and realign their sights on the target between shots.

Breathing Sensor

The Zephyr Technology BioHarness BT is a compact electronics module that attaches to a lightweight fabric strap with embedded sensors that monitor electrocardiogram signals and breathing waveforms. The sensor is worn like a heart rate monitor and is completely wireless. The device was used to monitor participants' breathing trends while firing the weapon.

Intelligent Tutoring System Framework

The Generalized Intelligent Framework for Tutoring (GIFT) is an evolving architecture-based project (Sottilare et al. 2012; access GIFT software through https://gifttutoring.org). This generalized approach enables system developers to quickly construct intelligent tutoring capabilities through a set of standardized tools and messaging schemas. For this study, GIFT provided the architecture required to create a unified marksmanship system from which data was collected and logged while subjects interacted with the FATS M100. Modules in GIFT were authored to enable communication protocols for the collection of the FATS M100 performance information, the FATS M100 sensor embedded rifle data, and metrics collected from the BioHarness wearable sensor. All data was logged on a synced time stamp, for easy merging in a post-hoc setting in support of model development and validation practices. With respect to this study, GIFT was primarily used as a data collection apparatus to sync data feeds from multiple channels onto a single logging timestamp to drive model development. How the models will be integrated and applied in GIFT is discussed below in the "Future Work" section.

Independent Measures

There are two independent variables (IVs) driving model development: firing stance and equipment setup (see Fig. 2). The firing stance IV consisted of two positions: (1) prone and (2) kneeling. The second IV, equipment setup, consisted of two variations of uniform/equipment arrangements: (1) wearing just the standard Army Combat Uniform (ACU) consisting of trouser, t-shirt and Army Combat Boot minus the blouse (i.e., camo); and (2) same as "camo" but wearing additional combat equipment consisting of helmet and outer tactical vest with Small Arms Protective Inserts (SAPI) (i.e., gear). As the goal of the study was to build models of expert performance, the IVs were used to distinguish marksmanship behaviors to determine if differences are observable across the two firing conditions and two gear setups.

Expert Model Attributes and Performance Parameters

Dependent measures are associated with two distinct categories. There are measures linked to performance outcomes that dictate the quality of a firing event, and there are measures linked to operator behaviors that occur during the execution of a shot. Individual shot behavior metrics were used to observe how an expert functions during a firing event and will be used to determine if experts consistently do the same things across each trial. To determine consistency in application, the procedure involves a shot group size. For this study, we apply a "five-shot group", which measures the distance between the two furthest shot points after an individual fires five shots while aiming at the center of target. The Shot Group Size, Aim Trace and Trigger Displacement measures were selected because the simulator already produces data supporting these metrics and because of their relationship to the functional elements of a shot

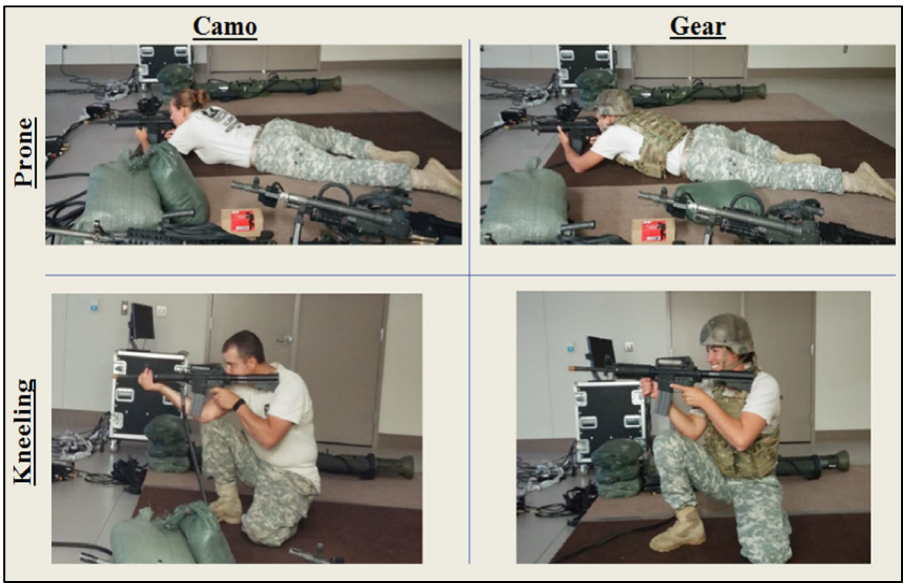

Fig. 2 Expert model development conditions

(Department of the Army 2016). Table 1 describes the measures being collected and their relationship in the process of expert model development.

In this effort we set out to model the behavior of expert marksmen in the act of using a standard issue M4 to place 5-shot groups on a 25 m zeroing target. In particular we sought to model per-shot behavior for three functional elements of marksmanship in temporal proximity to the fire event. The end goal was to develop a unified set of independent but complementary models against which the behavior of novices executing shots under comparable conditions could be used for assessment purposes in an ITS. The shooting activity was patterned after the "5-shot group" procedure specified in TC 3-9.22 for zeroing during initial marksmanship training (Department of the Army 2016). Though our focus was on modeling individual shot behavior, we considered group size as an aggregate overall performance measure, choosing to exclude from our modeling calculations shots belonging to groups with size greater than 4.0 cm, the performance requirement specified by the TC.

Procedure

Before participants arrived, the simulated M4 carbines were mechanically and digitally zeroed by the research team. This involves ensuring that the weapon sights are set so a shooter will not have to take ballistics into account when firing at the 25 m zeroing target. The simulated bullet strike will be exactly where they aimed. To confirm this, the researchers would then fire a few test five-shot groupings.

Upon arrival, participants received a brief overview of the study and were asked to fill out an informed consent form. Next, subjects were fitted with the BioHarness breathing strap, and the sensors were then synced to GIFT for time-synced data logging. Participants then filled out a Demographics Questionnaire covering experience in the AMU and the various awards they won throughout their careers.

Following the administrative portion of the experiment, participants were presented a short PowerPoint slideshow that reviewed the purpose of the study. They then received information on the FATS M100 simulator and all the sensors that collected data during task execution. Next, participants were given the opportunity for familiarization training with the system. They were instructed on tasks procedures, and then practiced with the weapon for 5 min on the 25-m range. During practice, the rifle was re-zeroed once for each individual participant. This was to account for their own particular setup, as the zero applied by the research team was based on a proctor's shot placement. The re-zero was also only applied during the practice portion, prior to main data collection window. The re-zero has no impact on the analysis, as the goal was to model shots across consistent performance and placement. Next, subjects were prepped for the main data collection window, where they were asked to fire at the center of the zeroing target at their own pace. They were informed that after every 5 shots fired (i.e., a "5-shot group") the carbine would become empty and that they would have to reload it with another 5-shot magazine. After each five-shot group, the participants were presented the placement of their shots and the associated group size score on screen.

Table 1 Performance and behavior metrics driving expert model development

Category	Measure type	Description	Data source	Expert model
Performance	Shot Group Size[a]	• Distance between the two shots (within a 5-shot group) that are furthest apart[b] • Used to gauge consistency across shots	FATS M100 Shot Group Algorithm	Used to define expert performance
Behavior	Breathing Waveform	• Real-time monitoring of respiration patterns during execution of a shot	BioHarness BT Strap	Breath Control
	Aim Trace	• Real-time capture of optical sensor readings in relation to the FATS M100	FATS M100 Optical Sensor	Body stability, control of weapon
	Trigger Displacement	• Real-time monitoring of the distance displacement of the carbine's trigger	FATS M100 Trigger Sensor	Trigger Control

[a] Size of group, regarded as a property of each shot within the group

[b] Independent of the group's position relative to the 4 cm circle on the 25 m zeroing target

All participants were given 18 min in each of the four conditions to fire as many 5-shot groups as they could with self-regulated breaks administered when needed to control the effects of fatigue. If a participant's shots were consistent in their location but not in the center of the target, the M4 was digitally zeroed again, simulating the mechanical process of a shooter on a real range adjusting their mechanical weapon sights to correct for individual sight picture differences. This re-zeroing process was applied only once for each participant, if necessary, to account for their particular set-up and eye relief. This did not alter the measurements collected but provided reassurance to the expert performers who were used to seeing their shots accurately represented. Condition interactions were sequenced in the same order for all participants; Camo-Prone, Camo-Kneeling, Gear-Prone, and Gear Kneeling. Upon completion of data collection with the FATS M100, participants completed a post-experiment survey.

For reference, the post-experiment survey was used to gauge reactions across each experiment as it related to the apparatus and their performance. All eight experts agreed they performed to the best of their ability, and zero experts claimed interference issues due to the sensorized and tethered carbine.

Data Capture and Analysis

In this section we describe the process involved in expert model development, including the capture and logging of data, initial data processing, and data reduction practices to produce variables to drive model creation.

Data Capture and Logging

Time-stamped data from all systems and sensors were recorded by GIFT. The FATS M-100 data were recorded as part of the domain session log using the GIFT logging protocol. Sensor data was logged directly by the sensor module. These logs were created by the system for each participant.

GIFT Plug-in for the Marksmanship Simulator and Breathing Sensor

A custom GIFT Gateway plug-in was developed to receive data from the Meggitt System over a TCP socket connection. Simple binary packet formats were established in collaboration with Meggitt staff.

Key packet types included:

- New Target
- Aim Data
- Weapon Sensor Data
- Fire Event
- Practice End

Aim data (based on x/y coordinates) and weapon sensor data were collected at 6 Hz. Time sensitive packets (aim, weapon sensor data, and fire events) were time stamped

by the Meggitt system. A time sync protocol was used to relate time stamps on the FATS M-100 with the GIFT domain session time.

A GIFT plug-in was developed for the Zephyr BioHarness to receive the data wirelessly via Bluetooth. Data types included ECG Waveform, Breathing Waveform, and a General Packet. The breathing waveform data was sampled at 18 Hz and was processed by GIFT in bundled packets.

Performance Criteria for Model Development

Before delving into the methodology applied to construct expert models, it is important to highlight the expertise of the sample from which the data was collected. As described above, each expert was asked to perform a series of five-shot groups across four conditions. The goal was to collect a large sample of experts doing what they do best, with performance criterion being defined to designate shots for inclusion in the model builds. A current standard threshold associated with rifle marksmanship grouping exercises is 4 cm for shot group size (Department of the Army 2016). Therefore, we deemed all 5-shot groups under 4 cm to be expert-quality shots for further analysis. For a visual breakdown of expert performance across all conditions, see Fig. 3.

In visually examining the performance outcomes across all experts, you can see the quality of shots as represented by standard mean, min, and max descriptive statistics. What is most noticeable is the variability in performance when comparing shots produced in the prone condition against those in the kneeling stance (namely, kneeling on average produces worse performance than the prone position). These outcomes display a homogenous representation for the prone conditions, while exhibiting a heterogeneous sample for the kneeling (see Table 2). This will impact the models developed, as experts with more qualifying groups will contribute significantly more to

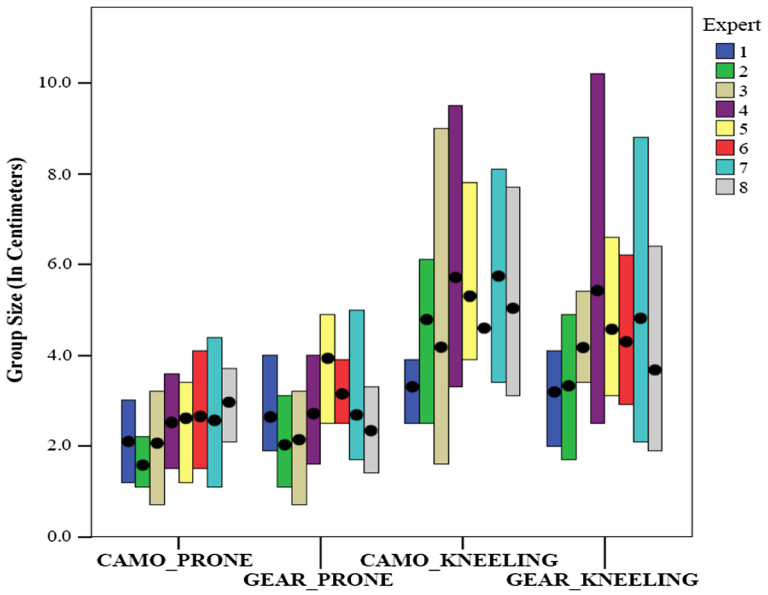

Fig. 3 Mean (black circle), min, and max shot group size for each expert in all associated conditions

Table 2 Number of 5-Shot groups per condition meeting the 4 cm performance criterion for each participant

Participant	Camo		Gear	
	Prone	Kneeling	Prone	Kneeling
1	10 (10)	7 (7)	8 (8)	6 (6)
2	10 (10)	2 (9)	11 (11)	7 (9)
3	12 (12)	7 (14)	17 (18)	6 (14)
4	11 (11)	2 (9)	9 (9)	3 (9)
5	10 (12)	*1 (7)	3 (10)	2 (8)
6	9 (9)	0 (1)	9 (9)	4 (9)
7	10 (14)	2 (10)	11 (12)	4 (12)
8	7 (10)	3 (8)	10 (10)	8 (13)
Average	9.9	4.0	9.8	5.0

Total number of groups fired are indicated within parentheses (*entry based on 4-shot group calculation due to dropped shot during data logging)

the resulting model. For a detailed statistical breakdown of performance comparisons across conditions, see Amburn et al. (2016). It is worth noting that we experienced technical difficulties with participant 6 in the camo/kneeling condition, where we only collected a single 5-shot group before a system crash. Due to time constraints we were unable to complete any further groups in that condition. As the produced group was above 4 cm, participant 6 behavior data was not accounted for in the model built for that condition. For reference, Table 2 lists the number of 5-shot groups collected for each participant in each of the four conditions that met our 4 cm performance criterion for inclusion on our model analyses.

Initial Data Processing

With a large sample of shots meeting the 4 cm performance criteria, the next phase in expert model development is examining each expert's behavioral characteristics across each shot. The first step in processing the data was to combine the various logs into a single CSV file containing all relevant experimental data using the GIFT Event Reporting Tool (ERT; Sinatra 2014). This tool synchronizes all data on a single source timestamp for appropriate temporal associations across all data streams collected at varying frequencies. After processing with the ERT, the data for each domain session was contained in a single CSV file having a row for each unique stamp. These files were then loaded into the Marksmanship Data Viewer (MDV, see Fig. 4), a tool developed specifically for viewing and manipulating marksmanship data exported from the GIFT ERT. The MDV has the ability to load multiple CSV files, allowing batch operations across multiple domain sessions.

After loading, domain sessions are presented to the user. Selecting a shot causes the MDV to display that shot's time-series data. Multiple shots can be selected and their time series overlaid for comparison. To facilitate the comparison of time series data from multiple shots the MDV adjusts the time scale of each shot so that $t = 0$ corresponds to the point where the trigger break occurred (i.e. when the shot was fired). In addition to manual selection, the MDV also supports shot selection using a filter. By entering filter criteria and then applying the filter, all shots matching the filter criteria will be selected (and overlays of their corresponding time series data

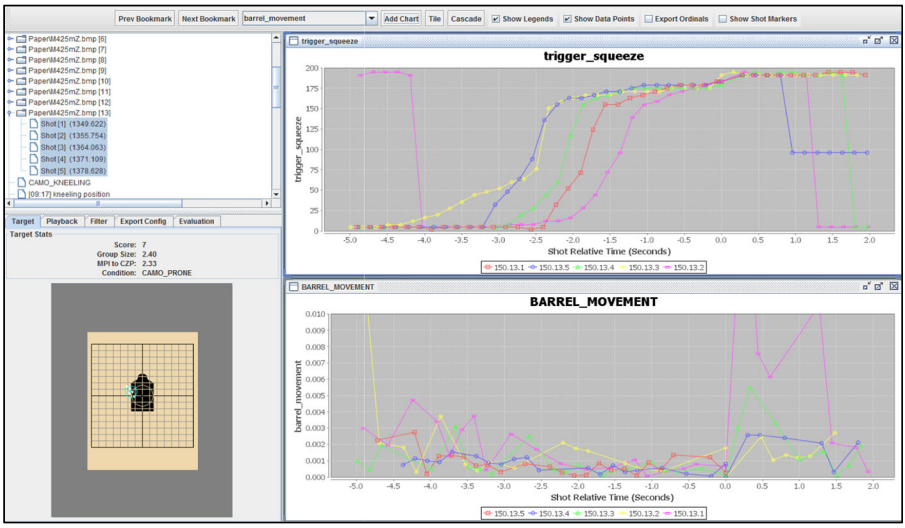

Fig. 4 GIFT's Marksmanship Data Viewer (MDV) with trigger squeeze and barrel movement data overlaid across a single 5-shot group for a participant in the camo-prone condition. (At time 0 is when the shot is fired, with trigger displacement at 175 units)

visualized). In this instance, we applied the filter to identify each shot associated with a 4 cm or less group size measure.

Time Series Data

The design of the system naturally generates several sets of time series data for each shot. These are referred to as raw time series data and include trigger displacement, aim X coordinate, aim Y coordinate, and breathing waveform. For analysis purposes, we found it useful to generate some derived time series as follows:

Barrel Movement Time Series To quantify barrel movement we combined the x and y coordinates of consecutive aim points using the distance formula:

$$distance_t = \sqrt{(\Delta x_t)^2 + (\Delta y_t)^2} \tag{1}$$

The aim points, as described under the data capture section, were derived from the aim data provided by the FATS M100, which is collected at a rate of 6 Hz. At each time step the distance is the distance (in normalized screen coordinates) that the aim point moved since the last time step. This derived variable provides information on the stability of an aim trace as it relates to distance between data points over a specified time window.

Trigger Squeeze Time Series The raw trigger squeeze data received from the Meggitt system was a value in the range of zero to 200, with values at or near zero indicative of no trigger displacement, and values near 200 indicative of full displacement. A

normalization procedure was performed by mapping the minimum signal value to 0.0 and the maximum signal value to 1.0. All other intermediate values were scaled to fit into range.

Breathing Waveform Time Series Firing during the natural respiratory pause in the breathing cycle is a common technique used during grouping (Department of the Army 2016). Consequently, our primary objective in analyzing the breathing waveform was to determine whether breathing was "quiet" during the time immediately preceding a given shot event. Observation of a live feed of the breathing waveform indicated that the signal reliably increases during an inhale, decreases during an exhale, and flattens out (albeit with some drift) when the breath is held. Furthermore, we observed from both live data as well as recorded data that the absolute value of the signal varies too much to be directly useful. For these reasons, we decided the best approach was to focus on the magnitude of the "deltas" (i.e., first derivative of the breathing waveform). Thus, at each time step the first derivative was computed by subtracting the signal value of the preceding time step.

Figure 5 shows an overlay of the breathing waveforms of each shot within a five-shot group. The increasing flatness of the curves as the time approaches the shot event ($t = 0$) is indicative of the shooter quieting their breathing for each shot. The trace for shot one (purple) is noticeably higher than the others, highlighting the need for a derivative approach rather than one based on absolute value.

Data Reduction and Qualitative Analysis

As mentioned previously, each shot generates several sets of time series data. So for each measure of marksmanship, we sought to distill the time series data for each shot

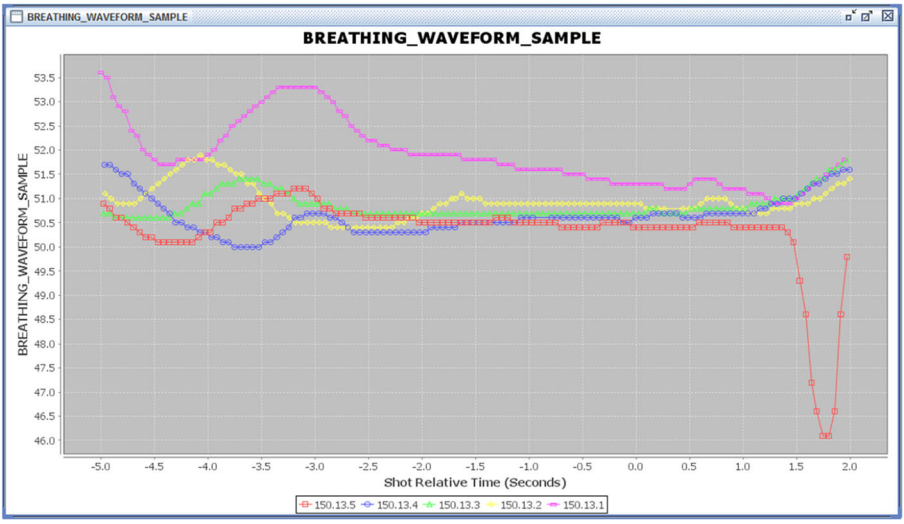

Fig. 5 Overlay of individual breathing waveforms for a sample 5-shot group for a participant in the camo-prone condition

into a single value reflecting the behavior on the relevant element. Doing so would provide two benefits: The first is that it would create an efficient way to measure each factor that can be assessed simply by comparing a shooters runtime value against a threshold value established by the expert representation. Secondly, it allows us to condense the multi-dimensional raw data set into manageable variable sets amenable to straightforward statistical analysis.

Subjective Interpretation of Expert Behavior

To implement this approach we started with subjective human interpretation guided by an understanding of the functional elements of marksmanship, as described in the U.S. Army TC (3-22.9; Department of the Army 2016). This analysis was reliant on the visualization capability of the MDV. In particular, we used the time-series overlay capability to compare and contrast numerous sets of shot data to identify noteworthy behavioral trends, especially in the period of time immediately surrounding a shot event.

After viewing just a few graph overlays several observations were made. First, the barrel movement (as evidenced by the aim trace data) invariably reduced to a minimum in the vicinity of the time the shot was fired. Secondly, as evidenced by the flatness of the breathing waveforms, the experts consistently quieted their breath leading up to the execution of shot, with the hold extending a rough half second following the release of the round. And lastly, shooters often started squeezing the trigger a second or more before the actual break of the trigger; the trigger was often held close to the breaking point (i.e., the point of resistance at which, if any more pressure is applied to the trigger, the weapon will fire) for a relatively long period of time before a shot was eventually fired (i.e., ranging between 750 and 2000 ms prior to shot execution).

Data Reduction to Represent Expert Behavior

Based on the subjective observations and the desire to reduce the behavioral data for each individual shot into a single value, a common statistical approach was applied for each model. This approach consisted of selecting a time interval (x-axis) in the vicinity of the trigger break and integrating the sensor signal (either raw or derived; y-axis) over that interval to calculate the area under the curve (AUC). As an example, we calculate the AUC for trigger squeeze on a normalized value of the signal for a time window of 1.5 s leading up to execution of the shot, resulting in a single value to describe the behavior of that signal for that time window. This descriptive metric provides a means for representing the behavioral trend of a variable over a window of time, with designated value thresholds serving as a means for performance classification. For both breathing and body stability (as inferred by barrel movement), a smaller AUC was viewed as desirable. This is in accordance with the accepted idea that minimizing barrel movement and quieting the breath during aiming and firing contribute to better and more consistent performance.

In the case of trigger squeeze, the desired behavior is for the trigger to be pulled straight back in a smooth motion. After analyzing the data, it was clear that the predominant behavior was an extended trigger squeeze, bringing the trigger near the breaking point and holding steady until finally applying enough additional squeeze to

fire the shot. Based upon this subjective observation, we chose to also perform an AUC calculation on this data feed; effectively quantifying the degree to which the shooter employed extended trigger control. To some extent the presence of an extended trigger squeeze almost guarantees that the first part of the trigger squeeze was smooth, because it is mechanically very difficult to pull the trigger to near the breaking point using anything other than a smooth motion.

Once the integration approach was established for all functional metrics, the final step was to select an integration interval and integration step size for each variable. The intervals were chosen based upon visual inspection of relevant time series data using the MDV. For trigger squeeze and barrel movement we chose the interval from 1.5 s prior to the shot, up until the shot was fired (i.e., from $t = -1.5$ to $t = 0.0$). For breathing, the interval also started at $t = -1.5$, but extended to t = +0.5; reflecting our observation that the experts typically kept their breath still for a half second after firing.

The requirements for the integration step size were two-fold. First we wanted the step size to divide evenly into the integration interval as a whole, and secondly we wanted it to be as small as or smaller than the interval between the incoming raw data points to take full advantage of the data. Since the weapon sensor (trigger and aim) data arrived at 6 Hz, we chose a 100 ms (0.100 s) step size for barrel movement and trigger. The breathing waveform data arrived very close to 18 Hz so we chose a 50 ms (0.050 s) step size for integrating the breathing derivative. For each integration step, the midpoint of the integration delta was calculated, and then signal value for that time was computed using linear interpolation between the nearest preceding and nearest succeeding data points.

With support for these computations (derived time series, integration approaches, and integration parameters) in place, we output the marksmanship data for all shots meeting our performance criteria to a single CSV file having one row per shot. The first several columns were set to contain data such as the domain session ID, target ID, shot ID, shooter-handedness, and the experimental condition (e.g., CAMO_PRONE | CAMO_KNEELING | GEAR_PRONE | GEAR_KNEELING). Additional columns were output according to the integration parameters described above, including AUC calculation for barrel movement, breathing, and trigger squeeze. This spreadsheet served as the basis for model development, where the AUC measurements were the inputs to build descriptive representations of behavior. With this data set in place, we performed a cross-fold validation test to determine the viability of a generalized expert model based on all observed shooters.

Results

In this section we define the model creation process and review the cross-fold validation method applied to evaluate the models diagnostic accuracy. Following, the results of a regression analysis are presented.

Model Creation with Cross-Fold Validation

An “n-1” leave-one-out cross-fold validation process was used to validate the accuracy of the developed expert models. The process was carried out as follows:

1. From the produced data set described above, set parameters to locate only shots associated with a 4 cm group size or smaller. The 4 cm group size threshold was used to identify shots that exhibited consistent behavioral application. An assumption with this approach is that a small group-size associates with consistent expert behavior, with the individually associated shots being ideal for modeling to determine the utility of a generalized model of fundamental application of marksmanship behaviors.
2. From this subset of expert performing shots, create a descriptive model of marksmen behavior using individual shot data from seven of the eight participants (the excluded participant will be used as a test case).
3. Resulting models (breathing, trigger control, and aim trace) are based on the mean of all AUC measures across all individual shots recorded within the shot groups meeting the performance criterion. Model thresholds are defined as any AUC value within two standard deviations of the mean to be classified as expert-like behavior.
4. Compare ("cross-check") the excluded participant's shots to the created model to determine whether the excluded participant exhibited expert-like behaviors; the behaviors in an accurate generalized model should align with the behaviors of any given expert (e.g., the excluded participant). A measure of how well the model performed is the percentage of shots by the excluded participant on which the participant's behaviors aligned with the model (within two standard deviations of the model mean or better).
5. Repeat the above process for each condition (stance, gear), behavior (barrel movement, breathing, and trigger control), and with each participant taking a turn as the excluded participant (used as the test case).

Cross-Fold Validation Results

Table 3 summarizes the results from one iteration of the aforementioned cross-fold validation process. A table of this nature was produced for each behavioral metrics (breathing, barrel movement and trigger control) and across each of the four firing conditions. This cross-check provides a means for identifying if there are existing statistical trends in behavioral application, and for recognizing outliers that produce expert-like results but perform the task in a dissimilar fashion to the others.

In evaluating model builds, it is important to identify extreme outliers, as they can impact models by skewing associated descriptive metrics. For instance, the trigger control model tracks trainees' behavior through the computed AUC measure, with steadier trigger pulls corresponding with larger values. In Table 3, one such outlier was identified: Participant 3's trigger control was recognized as being so unlike the other experts' (even with overall performance being very good) that the "mean minus 2SD" column is filled with negative numbers whenever Participant 3 is included in the model (negative values being an impossibility for AUC calculations). As a result, all shots when compared to models incorporating this participant's data were classified as expert. For this reason, we opted to create trigger control models with participant 3's data removed entirely. Table 4 summarizes the trigger control model tests with this omission.

Table 3 Cross-fold validation summary table for trigger control in the Gear/Prone condition

Excluded from model (used as test case)	Observation count			Trigger control model (higher = steadier)		Cross-check of excluded participant	
Participant	Excluded shots[a]	Included shots[b]	Total shots	Mean	Mean minus 2SD	Expert-like shots (trigger control)[c]	Pct shots like expert[c]
1	40	350	390	0.913	−0.0081	40	100%
2	55	335	390	0.951	−0.058	55	100%
3	85	305	390	1.157	0.583	9	11%
4	45	345	390	0.945	−0.066	45	100%
5	15	375	390	0.936	−0.042	15	100%
6	45	345	390	0.911	−0.083	45	100%
7	55	335	390	0.920	−0.096	55	100%
8	50	340	390	0.890	−0.089	50	100%

[a] Shots taken by excluded participant, which is the one indicated in the left-most column (the number of shots taken by each participant varies as a function of the number of 5-shot groups that met performance criteria)

[b] Shots meeting performance criteria taken by all remaining participants for that n-1 combination

[c] Comparing shots taken by the excluded participant (indicated in the left-most column) to the expert model built via the behaviors of the other seven participants and the associated 2SD threshold

With the goal of establishing generalized models of behavior strategy, excluding individuals who exhibit a different strategy is important to ensuring the generated model depicts the strategy being modeled. This assumption is supported by Siegler (1987, 1988) where strategies across different cohorts of students when learning math were identified and modeled, with exclusion of student data when their strategy was classified as different from the rest. When the trigger control models were built without Participant 3, the AUC values were much more homogeneous and representative of how most experts pull the trigger steadily; relative to the other experts, Participant 3

Table 4 Cross-fold validation table for trigger control in the Gear/Prone condition with participant 3 removed

Excluded from model (used as test case)	Observation count			Trigger control model (higher = steadier)		Cross-check of excluded participant	
Participant	Excluded shots	Included shots	Total shots	Mean	Mean minus 2SD	Expert-like shots (trigger control)	Pct shots like expert
1	40	265	305	1.139	0.542	40	100%
2	55	250	305	1.203	0.708	37	67%
3	–	–	–	–	–	–	–
4	45	360	305	1.186	0.633	40	89%
5	15	290	305	1.150	0.569	15	100%
6	45	260	305	1.141	0.553	43	96%
7	55	250	305	1.162	0.573	54	98%
8	50	255	305	1.118	0.534	49	98%

was "slapping" the trigger (i.e., pulling and releasing very quickly). As a result, the model's threshold for a participant being considered an expert is a bit higher in Table 4 than in Table 3, and Table 4 therefore served as a better threshold to compare experts against. For all conditions and all measures, participant 3 in trigger control was the only recognized outlier that warranted removal from the model builds.

These generated tables were then evaluated to determine the efficacy of a generalized behavioral expert model, and to further identify experts that exhibit different behaviors when performing marksmanship tasks. As an example, in Table 4, Participant 2 produced 55 qualifying shots, but only 67% of those shots exhibited expert-like trigger control when compared to their peers. In this instance, the model generated excluding this individual's data is deemed to be the most representative of proper trigger control from an assessment standpoint.

From our initial results, the data supports a generalized expert modeling approach that accounts for multiple behavioral representations on a single measure. In Table 5 we present an overall summary of all cross-fold validation results for all condition/metric pairings. In this layout, statistics are presented regarding the number of shots meeting performance criteria for inclusion in the model builds for each condition, along with the number of experts falling within performance categories based on percentage of shots meeting the 2-SD threshold during their cross-check procedure. For an expert's data to be included in that condition's generalized model for, we required each individual to have their behavior AUC values classify, at a minimum, 90% across their individual shots as expert-level. Anecdotally, for each measure and in each condition, at least one expert (but never more than two) exhibited significantly different behavioral techniques when evaluated through cross-check procedures. While the performance from these individuals met expert criteria, the behaviors recorded during execution of those shots were statistically different from their counterparts, and varied from the doctrinal descriptions of fundamental application. For this reason, those individuals were excluded from the condition model builds on the associated behavior classified as non-expert.

Regression Analysis

Following model creation and cross-fold validation checks, using SPSS Statistics 19, a final analysis on the expert data set was run to evaluate the influence these behavioral measures had on performance outcomes. The goal was to determine the best combination of behavioral variables for predicting marksmanship scores. For this purpose, we generated a performance value for each individual shot as it related to the shot group it was fired within. With shot group size being applied for determining consistency of marksmanship application, we calculate individual shot performance by measuring the distance of each shot location to the designated center of that cluster of five shots. In doing so, we can identify the shots within a group that had the greatest impact on overall group size calculations, along with being able to link each behavioral representation with a performance score that associates with its accuracy within that designated cluster.

With this new performance metric at the individual shot level, a stepwise regression test was performed on the entire data set (see Table 6 for results), with the individual

Table 5 Cross-fold validation summary table for all AUC measures across all firing conditions

	Number of excluded shots per participant (for use as test case)			Number of experts who had X% of shots correspond with behavior model								
	Min	Average	Max	Breathing			Barrel movement			Trigger control		
				90–100 (100)	80–89	<80	90–100 (100)	80–89	<80	90–100 (100)	80–89	<80
Camo/Prone	35[c]	49.4	60	7 (5)	1	0	7 (3)	1	0	7 (4)	0	1
Gear/Prone	15	48.8	85	5 (3)	2[a]	1	7 (2)	1	0	6 (2)	0	1
Camo/Kneeling	0	14.9	35	5 (5)	1	1[b]	6 (4)	0	1[b]	6 (6)	0	1[b]
Gear/Kneeling	10	25.0	40	7 (5)	0	1	6 (3)	1	1	6 (7)	0	1

Numbers in parentheses associate with number of experts classifying 100% of shots as expert

[a] Two participants' breathing patterns were classified as expert across 89% of shots. Based on this, these subject's data were included in the model builds

[b] In the Camo/Kneeling condition, one expert was removed due to system errors and no shot groups meeting performance thresholds. As such, the resulting model was based on six experts, as one was found to not meet cross-fold validation criteria standards

[c] The value of 35 is for Barrel Movement and Trigger Control. For one subject in the camo-prone condition, breathing data was not properly logged for a total of 12-shots, thus making the min value 23

Table 6 Results of the multiple stepwise regression analysis

	T	*P*	B	*F*	*df*	*P*	ΔR^2
Step 1							
Overall model				295.939	1, 1554	< .0001	.159
Firing stance	17.203	< .0001	.866				
Step 2							
Overall model				188.923	2, 1553	< .0001	.195
Firing stance	9.620	< .0001	.579				
Barrel movement	8.304	< .0001	156.668				
Step 3							
Overall model				171.142	3, 1552	<.0001	.202
Firing stance	9.632	< .0001	.577				
Barrel movement	8.653	< .0001	163.047				
Trigger control	3.989	< .0001	.224				

shot performance score defined as the dependent variable and all behavioral data and firing condition information entered in as predictors. At step 3 of the analysis the behaviors of barrel movement and trigger, as represented by the AUC value, and the stance from which an expert shot from (prone vs. kneeling) entered into the regression equation and were found to be significantly related to shot group size performance. The multiple correlation coefficient was .452, indicating that approximately 20% of the variance of an individual shot score could be accounted for by the stance in which a shooter fires from, and the control that individual had over their weapon's aiming and trigger. The gear setup and breathing AUC factors were not found to significantly impact performance and were therefore not entered into the regression equation; they were instead listed as exclusion variables (gear setup, $t(1555) = -1.326$, $p > .1$; breathing control, $t\ (1555) = -1.444$, $p > .1$).

Discussion

In an effort to develop an adaptive marksmanship training capability, the above results support the utility of a generalized model of expert performance for the purpose of diagnosing novice errors across fundamental principles within rifle marksmanship. The analysis is insightful as it aggregates functional elements of a psychomotor task into a set of unified conceptual models for driving coaching practices within a grouping exercise. Through cross-fold validation techniques, trends across experts when performing standard shot grouping procedures were identified based on AUC metrics. These AUC values, and their designated thresholds identified during model analyses, will serve as the basis for the assessment logic in the first closed-loop adaptive marksmanship tutoring system testbed. The outcomes supporting a generalized representation of expert behavior are reinforced in literature centered on the modeling of expert behaviors in cognitive problem domains. Siegler (1987, 1988) found experienced individuals in the domain of mathematics to apply consistently similar behavioral

techniques to solve problem sets, with analogous findings from Rauterberg (1995) in the domain of electronic data processing. Rautenberg posits that an expert's more complex cognitive mental representation of a problem space leads to less complex behavioral solutions when comparing expert to novice solution paths in that domain of study. While these conclusions are in support of a generalized expert model, the application of this approach in a psychomotor domain requires further validation to confirm the methods applied produce the diagnostic sensitivity required to inform relevant assessments for driving coaching and feedback.

What's interesting is a possible contradiction between what is described in the TC, and what is observed in the data. In particular for trigger control, we observe a slight discrepancy in what is described in the training doctrine and what was observed in the data. While the TC focuses on a smooth application of pressure through the shot, we observed the experts slowly remove the slack in the trigger, with an extended pause at the breakpoint prior to shot execution. This highlights a need to further apply modeling techniques to support a bottom-up validation of the task descriptions provided in the TC. These analyses could change the way doctrine is formalized with data-driven methods.

In integrating these models into GIFT's architecture, design decisions must be addressed that direct coaching decisions as derived from GIFT's Learning Effect Model (LEM; Sottilare 2015). In this context, GIFT must be able to assess both performance and behavior, and use this information to select a concept to remediate and a feedback strategy to intervene with. Therefore, creating logic to de-conflict performance and behavioral outcomes is required, along with building logic to de-conflict the recognition of multiple erroneous behaviors within a single shot instance. To drive these decisions, findings from the model analyses, along with results from the regression tests, are used to establish pedagogical production rules for deciding what concept to coach and when to coach it.

Handling the "Non-Expert" Experts

Before addressing pedagogical considerations based on model outcomes, it is important to discuss the nuances between performance and behavioral application in a psychomotor domain. In many instances, individuals apply unconventional methods to attain performance goals, with some attaining expert performance marks. In this research we found across all measures, and within each stance/gear combination, there was always one recognized expert who performed at an optimal level while exhibiting statistically different behaviors as read by the sensing technologies. In terms of pedagogical management, this type of performer must be accounted for. From a behavioral standpoint, this individual would be recognized as erroneously performing fundamental behaviors, while registering expert-like performance outcomes. These competing assessments must be resolved to prescribe the most appropriate feedback for that trainee. As an example, if a tutor was developed to train overhand free throw technique in the domain of basketball, how would the system operate for a player who shot very well underhanded? Naturally, the expert behavioral models will determine that the performer is not executing proper technique as deemed by a traditional overhand model, yet their performance could register accuracy outcomes equal to or better than that of an expert.

Should an ITS intervene and correct that individual's behavior, or should it allow the individual to continue with their technique as long as performance remains above threshold? Based on this, we believe performance should be the overarching determinant of feedback types. Only when performance is not optimal should behavioral models be applied to determine the cause for erroneous application. Yet, for some individuals, subtle feedback on identified non-expert behaviors while producing expert-like performance might be enough for that individual to perform even better. Accordingly, future research is required to better determine how to handle these unique relationships and to optimize learning interventions based on empirical evidence.

Insights from Regression Analysis

In further breaking down the role of expert models in assessing functional elements of rifle marksmanship, it is important to understand the relationship between each element and performance. In addition, it is also important to determine if the assessed independent variables produced enough variations in behavior to constitute developing separate assessment models across both stance and gear conditions (see Amburn et al. 2016 for additional analyses on gear effects). To examine these relationships, a stepwise multiple regression was performed.

The outcome of this multiple regression identified firing stance as the most predictive behavior to shot group size. This variable alone accounted for 16% of the variance in performance, with barrel movement and trigger control also being identified as contributing variables to the regression equation (see Table 6). With barrel movement shown as a significant predictor, existing aim trace feedback tools already built into the EST can be applied for coaching purposes, but their effect must be studied.

However, the most interesting finding from this test is the identification of both breathing AUC and presence of gear as exclusion variables to the model (i.e., they were not found to significantly contribute to the model's predictive power). This is significant, as this study highlights the function of breath control to have little to no influence over an individual's resulting performance across shots (see Fig. 6). As you can see in the graphic, there are many instances where non-steady breath patterns (signified by higher value AUC measures) result in expert performance, and there are instances where a steady quiet breath result in poor performance outcomes, showing little correlation across the expert population. Furthermore, when comparing gear effects across both prone and kneeling, those in gear exhibited a slightly noisier breath pattern than those wearing just camo. Performance outcomes showed no difference within the prone position; however, despite not breathing as steadily with gear in the kneeling position, the experts on average produced smaller group sizes (see Amburn et al. 2016 for full statistical breakdown). This finding de-emphasizes breathing's effect on shot accuracy, and is in contrast to the resulting regression equations reported by Nagashima et al. (2008), where they found breathing and trigger control as significant predictors in determining novice from expert. However, it is difficult to compare the two studies as this approach examines expert behavior alone and its impact on performance outcomes. For a further breakdown of comparisons across experimental conditions, see Amburn et al. (2016). With plans for novice data collection in place, future analyses will be conducted to confirm these findings, and to establish profiles to better inform assessment practices. It is important to state that these claims are based on the expert

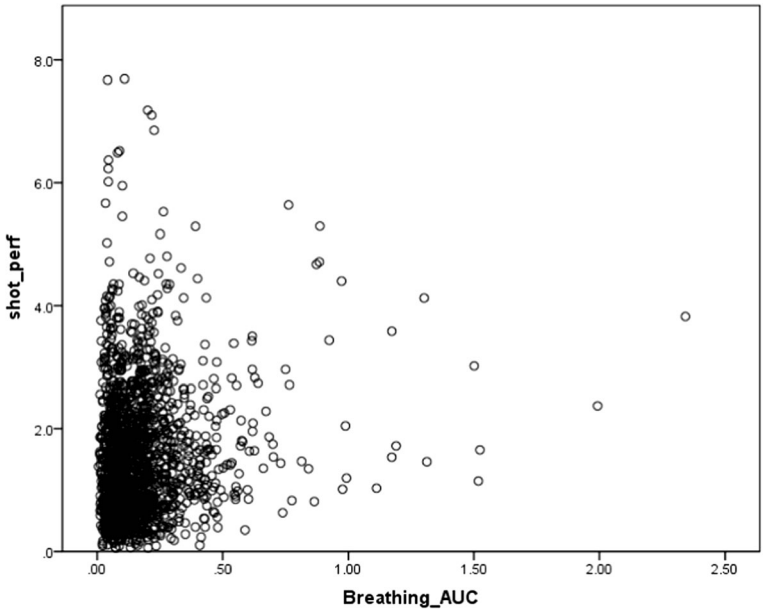

Fig. 6 Plot of breathing AUC values in relation to single-shot performance as computed for the regression analysis

performers from whose data our analysis was conducted. Further research is required to investigate the role breathing has on novice and journeyman performance.

Initially, the outcomes of this analysis will be applied for pedagogical purposes. The pedagogy should be considered from two perspectives: (1) in a self-regulated training environment where GIFT performs all assessment and provides all coaching, and (2) in an instructor facilitated event where GIFT provides support to the instructor and coaching to the trainee when appropriate. For the self-regulated use case, pedagogical considerations are made based on the assessment practices in place. It is not the ideal situation, as the models do not account for the critical functional element of the aiming process. However, to provide an initial coaching capability based on available data sources, some inference procedures needed to be defined. In the instance where all functional element behavioral measures are reporting as non-expert, we will apply the outcomes from this analysis to rank the concepts for conflict resolution purposes. According to the model, the most important component to correct first is a stable position to achieve steady barrel movement, followed by coaching a proper trigger control technique, followed by instruction on breathing effects during shots. While we cannot directly assess how an individual is aiming during a shot, we apply a pedagogical approach that targets a single functional element at a time. The goal is to target the functional elements we can monitor and track, with focused procedural training applied for the aiming process when consistent behavior is recorded across the remaining shot behaviors. In the second use case with an instructor in the loop, the pedagogical approach can be modified, as GIFT can be applied as a decision support tool. In this instance, we can use the instructor as an additional assessment resource, where they can apply their subjective interpretation, including elements related to aiming and sight alignment.

Future Work

With established expert models holding up to cross-fold validation procedures, the next step is integrating the model features in GIFT to establish a fully automated closed-loop testbed, which requires two primary development tasks. First, GIFT must be able to perform, in real time, the statistical computations the models are based on. With the ability to consume and log all behavioral information already in place, condition classes must be established for computing AUC values based on preconfigured parameters (e.g., transform trigger signal data into normalized range from 0 to 1 and compute AUC for time interval $t = -1.5$ s to $t = 0$ s). With this capability, GIFT can be configured to act on AUC values for triggering feedback interventions, which requires establishing domain representations within GIFT's authoring schema that adheres to the domain-independent nature of all remaining modules.

In GIFT, the authoring schema applied for managing real-time assessment in external applications is GIFT's Domain Knowledge File (DKF). When building a DKF, the author is responsible for three elements: (1) building an ontological hierarchy of concepts a given training event will assess (e.g., the functional elements of rifle marksmanship), (2) building assessment logic across all identified concepts based on available data and established condition classes, and (3) building instructional interventions (i.e., coaching prompts) that will be triggered based on assessment outcomes.

In building the assessment logic, the author has the ability to define three levels of performance based on GIFT standards (below-, at-, and above-expectation performance state messages; see Table 7 for defined thresholds in GIFT's DKF). Each of these expectation classifications have their own designated thresholds for this marksmanship testbed, where an AUC within 1 standard deviation of the expert mean is designated above-expectation and an AUC within 1 SD and 2 SD are labeled at-expectation. After each shot, these associated performance states are determined and logged. After a completed 5-shot group, each functional element produces a score based on the DKF assessments (above-expectation is worth two points, at-expectation is worth one, and below-expectation is worth zero). If a group size is larger than the designated 4 cm threshold, the behavioral functional element with the lowest score is selected for remediation. In the event of a tie, the highest ranked concept based on the regression analysis will be selected. It is also worth noting that our assessment techniques only cover three of the four functional elements targeted for this study. With no viable source to provide data on how an individual is aiming their rifle (i.e., aligning the front sight post over a target's center of mass, focusing their eyes on the front sight post and not the target, etc.), we apply an inference procedure to manage this elements' assessment. If all behaviors register within expert thresholds (steady barrel movement, steady trigger control, and a quiet breath) but their performance is above 4 cm, we infer that individual is varying their sight picture from shot to shot. This is consistent with Army doctrine which states, "When both a solid position and a good trigger squeeze are achieved, any induced shooting errors can be attributed to the aiming process for refinement" (Department of the Army 2016, p. 8–2).

When authoring instructional interventions, the developer can select what a trainee will see if a fundamental is selected for remediation. The author also has the ability to enter multiple variations of feedback and interventions, allowing for escalating levels if a concept is erroneously performed in multiple instances. These configurations are also

Table 7 Expert model thresholds configured in GIFT's DKF assessment schema where values represent calculated AUC measures as described above in the Data Reduction subsection (All measures outside of at-expectation range are classified as below-expectation)

Barrel movement		Breathing		Trigger squeeze	
Interval PreShot	−1.5 Sec	Interval PreShot	−1.5 sec	Interval PreShot	−1.5 sec
Interval PostShot	0.0 Sec	Interval PostShot	0.5 sec	Interval PostShot	0.0 sec
Integrate Interval	0.1 Sec	Integrate Interval	0.05 sec	Integrate Interval	0.1 sec
Stance Type: CAMO Kneeling		Stance Type: CAMO Kneeling		Stance Type: CAMO Kneeling	
Above Expectation	0.00392 AUC	Above Expectation	0.189 AUC	Above Expectation	0.957 AUC
At Expectation	0.00519 AUC	At Expectation	0.284 AUC	At Expectation	0.695 AUC
Stance Type: CAMO Prone		Stance Type: CAMO Prone		Stance Type: CAMO Prone	
Above Expectation	0.00165 AUC	Above Expectation	0.219 AUC	Above Expectation	0.960 AUC
At Expectation	0.00226 AUC	At Expectation	0.321 AUC	At Expectation	0.712 AUC

entered in the DKF schema and link feedback content to instructional requests from GIFT's pedagogical module. For the initial coaching system using the generated models, three levels of feedback are configured for each functional element being assessed. The first level of feedback is a single summary slide highlighting important aspects of the functional element being coached. The second level consists of a short video displaying an AMU instructor explaining the functional element being coached. And the third level is a detailed slide deck breaking down the functional element into its piece parts. This approach is being applied to support the first closed-loop testbed and will be used for an initial experiment with novice participants receiving ITS led marksmanship training.

It is important to discuss the implications of using an expert modeling approach to drive the pedagogical decisions described above. While the models derived from the AMU experts can be applied to determine if an individual is or is not behaving like an expert, the models do not have the diagnostic power to determine what error is occurring when expert behavior is not observed. In these instances the pedagogical functions are limited to coaching interventions on the techniques the models are based around. To explore extending the assessment logic even further, the notion of a buggy-library modeling approach is being considered (Pavlik et al. 2013). The challenge is generating a labeled data set that accurately and consistently classifies common novice behaviors from expert annotation. To support this modeling approach, we are generating a labeled data set of shots based on assessments performed by an expert human instructor during a novice data collection. Post-hoc analyses will allow us to determine if shots classified with the same annotated label contain statistically similar behavioral representations to build assessment logic around. This buggy-library approach would enhance the diagnostic power of GIFT assessments to account for common novice misconceptions that would ultimately inform more prescriptive feedback content.

Motor Response Inter-Dependency

Our analysis treated the three behavioral measures independently, yet they are clearly not independent. Most notably, we used barrel movement as a proxy for a stable body position, however barrel movement can be adversely impacted by both poor breathing technique and poor trigger squeeze technique. No doubt, breath and trigger skills are taught precisely because of their beneficial effect on barrel stability and the resulting improvement in aim quality.

Unraveling the interdependencies of these variables could be crucial for the eventual creation of an effective automated adaptive marksmanship tutoring system. A clear understanding of these relationships could lead to optimization of instruction via prioritized sequencing of targeted pedagogy. For the current effort our intent was simply to capture and characterize the behavior of expert shooters while in the act of creating quality shot groups. Interdependent or not, the captured data reflects how the expert shooters perform under the conditions set forth. With appropriate statistical analysis of the captured data we could conceivably shed some light on the various interdependencies; however, such an investigation would be more revealing if performed on a data captured from

cohort of shooters representing a broad range of marksmanship abilities. Identifying causal linkages can improve the assessment practices by demonstrating the impact specific behaviors have on shooter stability, and as a result, improve pedagogical practices that adhere to procedural coaching for promoting effective behavioral application. Nevertheless this is certainly a promising avenue worthy of future exploration.

Extending Methods Beyond Marksmanship Grouping Fundamentals

The initial scope of the first adaptive marksmanship capability is focused on the skills linked to grouping exercises (i.e., getting an individual to consistently strike a target in a repeatable fashion), and the models presented above fit that context alone. Further application of ITS methods to more advanced marksmanship skill sets (e.g., hitting targets with limited exposure times at varying distances, hitting moving targets, etc.) requires additional research to determine the best way to model the task's behavioral variables that would drive assessment practices. The models created are limited to the environment they were established within, and should not extend to training events outside basic rifle marksmanship task types.

Conclusion

This work is among the first to address the need for automated, personalized, and intelligent psychomotor training. Methods for measurement, model creation, and expert validation were addressed, along with their implications of being implemented in GIFT's domain-independent architecture. In terms of modeling psychomotor domains to guide adaptive instructional methods, this work lends itself to lessons learned that can inform recommendations for future practitioners and future developmental features native to GIFT's authoring and run-time environments.

In terms of furthering the work presented, the next phase is to run a validation study using novices with little to no experience handling rifles. Through this approach we can evaluate the efficacy of using the AUC thresholds to diagnose errors in accordance with some of the functional elements of rifle marksmanship, as well as determine if feedback generated from these assessments can ultimately improve performance over time. During the writing of this paper, we are currently analyzing data collected from a four-condition experimental design: (1) ITS managed assessment and feedback delivery, (2) randomized feedback delivery regardless of assessment outcomes (used to determine if assessment driven feedback is better than feedback by itself), (3) a human coach condition with a rifle marksmanship instructor (ideal case), and (4) a no-feedback control condition. This study will provide initial training effectiveness measures on the methods applied and will inform modifications to the training materials used for coaching.

References

Amburn, C., Goldberg, B., Brawner, K. (2014). Steps towards adaptive psychomotor instruction. In *Proceedings of Twenty-Seventh International Florida Artificial Intelligence Research Society (FLAIRS) Conference*. Pensacola Beach.

Amburn, C., Goldberg, B., Chen, D., Ragusa, C., Boyce, M., & Shorter, P. (2016). Effects of equipment on model development for adaptive marksmanship trainers. *In Proceedings of the 2016 Interservice/Industry Training Simulation and Education Conference (I/ITSEC)*. Orlando, FL.

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/10.3102/0013189X013006004.

Chung, G. K., Nagashima, S. O., Espinosa, P. D., Berka, C., & Baker, E. L. (2009). *An exploratory investigation of the effect of individualized computer-based instruction on rifle marksmanship performance and skill (CRESST Report 754)*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Department of the Army (2011). The U.S. Army learning concept for 2015. Fort Eustis: Training and Doctrine Command. Retrieved from http://www-tradoc.army.mil/tpubs/pams/tp525-8-2.pdf.

Department of the Army. (2016). *Rifle and carbine training circular (TC 3-22.9)*. Washington, D.C.: Army Publishing Directorate.

Goldberg, B. (2016). Intelligent tutoring gets physical: Coaching the physical learner by modeling the physical world. In *Proceedings of 2016 International Conference on Foundations of Augmented Cognition*. Toronto.

Goldberg, B., Sottilare, R., Brawner, K., Sinatra, A., Ososky, S. (Eds.) (2015). Developing a generalized intelligent framework for tutoring (GIFT): Informing design through a community of practice. In *Workshop Proceedings at the 2015 International Conference on Artificial Intelligence in Education (AIED)*. Madrid.

James, D. R., & Dyer, J. L. (2011). *Rifle marksmanship diagnostic and training guide (ARI Research Product 2011-07)*. Arlington: U.S. Army Research Institute for the Behavioral and Social Sciences (DTIC ADA544533).

Kulik, J. A., & Fletcher, J. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78.

Lieberman, J., & Breazeal, C. (2007). TIKL: Development of a wearable vibrotactile feedback suit for improved human motor learning. *IEEE Transactions on Robotics, 23*(5), 919–926.

Meggitt Training Systems (n.d.). FATS 100e Advanced Reality Training Simulator. Retrieved from: https://www.meggitttrainingsystems.com/Military/Simulation-training/FATS-100e-advanced-reality-training-simulator.

Nagashima, S. O., Chung, G. K. W. K., Espinosa, P. D., Berka, C., & Baker, E. L. (2008). *Assessment of rifle marksmanship skill using sensor-based measures*. Los Angeles: University of California, National Center for research on evaluation, standards, and student testing (CRESST).

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods*. New York: Wiley.

Pavlik Jr, P. I., Brawner, K., Olney, A., & Mitrovic, A. (2013). A review of student models used in intelligent tutoring systems. *Design recommendations for intelligent tutoring systems: Volume 1-learner modeling*. Aberdeen Proving Grounds: U.S. Army Research Laboratory.

Platte, W. L., & Powers, J. J. (2008). *Using motion capture to determine marksmanship shooting profiles: Teaching soldiers to shoot better faster*. Monterey: Naval Postgraduate School (DTIC ADA488985).

Pojman, N., Behneman, A., Kintz, N., Johnson, R., Chung, G., Nagashima, S., ... Berka, C. (2009). Characterizing the psychophysiological profile of expert and novice marksmen. In *Proceedings of 2009 International conference on foundations of augmented cognition*. San Diego.

Ranes, B., Lawson, B., King, M., & Dailey, J. (2014). *Effects of rifle handling, target acquisition, and trigger control on simulated shooting performance. (USAARL Report No. 2014-19)*. Fort Rucker: U.S. Army Aeromedical Research Laboratory (DTIC ADA601359).

Rauterberg, M. (1995). About faults, errors, and other dangerous things. In H. Stassen & P. Wieringa (Eds.), *Proceedings of the XIV European Annual Conference on Human Decision Making and Manual Control*. Delft: Delft University of Technology Faculty of Mechanical Engineering and Marine Technology.

Rickel, J., & Johnson, W. L. (1999). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence, 13*(4–5), 343–382.

Ritter, F., & Feurzeig, W. (1988). Teaching real-time tactical thinking. In J. Psotka, L. D. Massey, & S. A. Mutter (Eds.), *Intelligent tutoring systems: Lessons learned* (pp. 285–301). Hillsdale: Lawrence Erlbaum Associates.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology, 115*, 250–264.

Siegler, R. S. (1988). Individual differences in strategy choices: Good students, not-so-good students, and perfectionists. *Child Development, 59*(4), 833–851.

Sinatra, A. M. (2014). The research psychologist's guide to GIFT. In *Proceedings of the 2nd Annual GIFT Users Symposium*. Pittsburgh.

Sottilare, R. A. (2015). Challenges in moving adaptive training & education from state-of-art to state-of-practice. *In Workshop Proceedings at the 2015 International Conference on Artificial Intelligence in Education (AIED)*. Madrid, Spain.

Sottilare, R., Goldberg, B., Brawner, K. W., Holden, H. (2012). Modular framework to support the authoring and assessment of adaptive computer-based tutoring systems. *In Proceedings of 2012 Interservice/ Industry Training, Simulation, and Education Conference (I/ITSEC)*. Orlando.

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology, 105*, 970–987. https://doi.org/10.1037/a0032447.

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology, 106*, 331–347. https://doi.org/10.1037/a0034752.

Stottler, D., Harmon, N., & Michalak, P. (2001). Transitioning an ITS developed for schoolhouse use to the fleet: Tao ITS, a case study. In *Proceedings of 2001 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. Orlando. (DTIC ADA597998).

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*, 197–221. https://doi.org/10.1080/00461520.2011.611369.

Yates, W. W. (2004). *A training transfer study of the indoor simulated marksmanship trainer*. Monterey: Naval Postgraduate School (DTIC ADA427059).