Intelligent Tutoring Authoring Tools to Assess Training Effectiveness

Rodney Long

Army Research Laboratory
Orlando, FL
Rodney.A.Long3.Civ@mail.mil

Mike Smith, Sue Dass, Clarence Dillon,
Julie Silverman
ICF International
Fairfax, VA
Mike.Smith@icfi.com, Sue.Dass@icfi.com,
Clarence.Dillon@icfi.com, Julie.Silverman@icfi.com

ABSTRACT

Intelligent Tutoring Systems (ITS) hold the potential to unlock a new era of adaptive, learner-centric training, but much of the current research focuses on learner-tutor interactions. Alternatively, this paper describes ongoing research to demonstrate an automated data analysis capability that supports training effectiveness evaluation from the tutor authoring perspective. The research team investigated: how to extend the Army Research Laboratory's Generalized Intelligent Framework for Tutoring (GIFT) to provide this automated analysis capability; what data collection mechanisms could be utilized to support the analysis; and how to present the outputs to support decision-making about instructional strategy in an ITS. The resulting proof-of-concept operationalized this framework to support the rapid, high-level, visually intuitive analysis of effectiveness at a user-selected level of granularity and to then offer a mechanism to delve deeper and explore individual factors to ultimately identify areas for improvement.

The research team simulated experimental learner demographic and performance data to verify and evaluate the proposed methods since there were no examples of completed GIFT courses at the time of this research. A statistical engine was used to identify factors that contribute significantly to training effectiveness and to support investigation of the research questions. An Army marksmanship course was chosen as a use case since it relies on multiple training delivery techniques and includes factors external to the GIFT environment, such as experience with first-person shooters or prior experience. The research utilized standard GIFT data sources alongside Experience Application Programming Interface (xAPI) formatted data to combine disparate sources and support integration of multiple perspectives. While the extent to which the experimental data represents real-world use cases will have to wait for the ongoing GIFT courses to be completed, the experimental data provided a foundation for initial technical evaluation of this proof-of-concept system based on existing data sources.

ABOUT THE AUTHORS

Rodney Long is a Science and Technology Manager at the Army Research Laboratory in Florida and is currently conducting research in adaptive training technologies. Mr. Long has a wide range of simulation and training experience spanning 28 years in the Department of Defense and has a Bachelor's Degree in Computer Engineering from the University of South Carolina and Master's degree in Industrial Engineering from the University of Central Florida.

Mike Smith is a Technical Specialist with ICF International and has over 12 years of experience in data analytics, strategic planning, and risk assessment. Mr. Smith currently advises several clients on how to adapt emerging analytics practices to improve their organizational performance. Mr. Smith has a B.A. in International Economics from Longwood University and a Master of Public Policy from Georgetown University and is a Certified Analytics Professional (accredited by the Institute for Operations Research and Management Science).

Sue Dass is a Technical Specialist at ICF International and has over 15 years researching, designing, developing, and managing instructional projects for private and government organizations using advanced learning technologies. Dr. Dass recently co-designed an electronic performance support tool to help faculty explore, select, and implement learning technologies. Dr. Dass has a B.S. in Civil Engineering, a M.Ed. in Instructional Design, and a Ph.D. in Education.

Clarence Dillon is a computational social scientist and Consultant at ICF International. He authored the ontology for the Department of Defense strategic planning scenarios and established the first collaborative, semantic web platform

used in DoD. Mr. Dillon holds a Bachelor of Arts in International Affairs, a Masters of Social Science in International Relations and he conducts graduate research in social complexity at George Mason University.

Julie Silverman is an Associate at ICF International focusing on business process improvement and strategic planning. Ms. Silverman works with federal clients to improve overall efficiency and workforce productivity through process improvement, leadership development, and organizational assessments. Ms. Silverman has a Bachelor’s of Business Administration in Marketing, with a dual major in History from the College of William and Mary.

Intelligent Tutoring Authoring Tools to Assess Training Effectiveness

Rodney Long

Army Research Laboratory
Orlando, FL
Rodney.A.Long3.Civ@mail.mil

Mike Smith, Sue Dass, Clarence Dillon,
Julie Silverman
ICF International
Fairfax, VA
Mike.Smith@icfi.com, Sue.Dass@icfi.com,
Clarence.Dillon@icfi.com, Julie.Silverman@icfi.com

INTRODUCTION

Intelligent Tutoring Systems (ITS) hold the potential to unlock a new era of adaptive learner-centric training, but much of the current research focuses on learner-tutor interactions. Alternatively, this paper describes ongoing research to support training effectiveness evaluations from the tutor authoring perspective. The Army Research Laboratory's Generalized Intelligent Framework for Tutoring (GIFT) is being extended to explore training effectiveness within an ITS. The researchers have developed a proof-of-concept tool for providing visually intuitive analysis of GIFT courses that also incorporates user-defined external data sources. The resulting platform provides authors with varying levels of statistical competency with an accessible, intuitive user experience that reflects best practices for course evaluation. The technical architecture was built on a flexible and extensible suite of technology that utilized several popular open source data science tools to provide a foundation for future growth and a model for other applications.

In contrast to other research on analytics dashboards or training effectiveness, this paper focuses on demonstrating the enhanced GIFT user experience and presents several examples in the results section of the paper. The methodology section describes the various components of the proof-of-concept and their role in the final product. The prior research section provides a brief overview of the supporting literature underpinning the design. The data generation section describes the simulations utilized to provide experimental data for analysis and the technology selected for the proof-of-concept. The paper closes with a future research section that touches on potential extensions for the next phase of research and touches on some of the non-technical hurdles to adoption of this and similar tools and technologies.

METHODOLOGY

The primary research objective was to demonstrate an automated data analysis capability that supports training effectiveness evaluation and performance interventions. To this end, the research team investigated how the existing GIFT architecture could be extended to provide this automated analysis capability, what data collection mechanisms could be utilized to support the analysis, and how to present the outputs to support decision-making about instructional strategy in an ITS. Since there were no examples of completed GIFT courses at the time of the research, the team simulated experimental learner demographic and performance data to verify that the system was capable of performing the analytical functions as intended and to provide a context to evaluate the research questions. The team drew from publicly available records on demographics, existing performance data from other sources, and existing research on the selected use case in order to develop an experimental data set that represents plausible real-world distributions and results. While this approach left larger research questions about usability or effectiveness in the author or learner context beyond the scope of this research, the experimental data provided a foundation for initial technical evaluation of this proof-of-concept system based on existing data sources within and external to GIFT.

While GIFT has embedded tools for collecting user profile, assessments, and other survey data, tools have yet to be developed that allow an author to evaluate instructional performance and make changes based on findings. Rather than focus on data analysis procedures themselves, this paper describes the development of a generalized platform that will form the basis for an automated analytics capability that can be applied to a broad array of course types and assessment strategies. The primary user perspective envisioned was a tutor author or administrator concerned with standards and evaluation of instructional content, with GIFT experimenters and researchers being secondary potential users.

While not explicitly necessary for this purpose, Basic Rifle Marksmanship (BRM) was selected as a use case to provide context, sample data, and a basis of research for evaluation. In the U.S. Army, all soldiers begin training with BRM, which focuses on the fundamentals of marksmanship and incorporates multiple modalities of course work to include: classroom, simulation, and live-fire training. This context provided a plethora of learning variables that could be observed, documented, and analyzed. These variables were sorted into three main categories: learner profile; physical measurements; and other external variables (see Table 1). Marksmanship performance also encompasses individual performance based on the results of a single course participant, as well as group performance based on a team, cohort, or platoon that could eventually be incorporated into the scope of the research. In addition, marksmanship performance assessment data was readily available in the Experience Application Programming Interface (xAPI) format, as described by Hruska, Amburn, Long, Kilcullen, and Poeppelman (2014).

Table 1. Selected Demographic Variables

Standard	Non-Standard
Age	Sport handgun proficiency
Gender	Sport hunting rifle proficiency
Educational level	Last fired a weapon
Years in active/reservist duty	Hours per month at shooting range
Functional area/job specialty	Formal marksmanship training
Years in functional area/job specialty	Perceived knowledge level
Location	Last formal course training
Deployed	Performance on last qualification
Months deployed	Gamer
Anxiety*	Hours per week first shooter game
Self-efficacy*	
Motivation*	

*Although standard, these constructs are likely more informative when customized to a performance measure

As discussed in detail in the Data Generation section, our research used a combination of recorded fire and simulated experimental data based on the components of the marksmanship course to show outcomes of the course and ostensibly make predictions based on student experience and course success. In order to support the initial technical evaluation of system functionality, the research was used to inform likely distributions of the sample data as well as dependencies that would likely be observed in real-world data. For example, students who entered the course as experienced marksman performed better on the observed results. While this approach does not constitute a field test of the proof-of-concept based on real-world observations, it did provide a plausible context for evaluation of the components and functions of the analytic system, as well as context to present actionable findings, and demonstrates how existing data collection mechanisms in GIFT could be utilized.

Rather than presenting an overwhelming array of analytic outputs to the user, the design used a layered approach to data visualization. Health indicators provide a visual representation of aggregated data sets that effectively present potential instructional issues, and allow the user to delve into additional details, as needed. This structured approach used a set of standard data items that could be made available for every GIFT course, though the proof of concept also leveraged external data to evaluate the marksmanship use case. Standard Data are defined as data sources that are currently available in the GIFT architecture through the survey functionality or through other student-tutor interactions and thus could potentially be utilized for any future course in GIFT. What are referred to throughout as Non-Standard Data items are contextual and were selected based on the marksmanship research. A mix of these two different data types are utilized to demonstrate analytic capabilities drawn from diverse data sources. Future capabilities will be added to allow for more open-ended analysis with a broad array of methods and visualizations and to support analysis of diverse data types, formats, and sources.

PRIOR RESEARCH

While the research drew on intelligent tutoring and marksmanship research to support its approach, the focus was on a generalizable evaluation framework that could be applicable to any instruction utilizing GIFT. There is already a depth of research on both topics that this effort did not seek to expand on directly. This research drew on the framework presented in Long, Smith, Dass, Dillon, and Hill (2016) that describes three use cases for analytics applications. Specifically, this research aimed to enhance the post evaluation use case by focusing on insights that can be developed both within and eventually across instructional content. In order to develop the dashboard, a literature review was conducted to provide insight into recent advances in intelligent tutoring, as well as methods and techniques for evaluation dashboards, and to determine the best variables to use for evaluating the marksmanship use case.

While the concept of intelligent tutoring is not new, recent research has focused on the efficacy of adaptive training and automated student-tutor interactions. Some adaptive offerings have emerged on the commercial market, with the most notable being the Knewton adaptive learning platform, whose website (https://www.knewton.com) contains a host of research purporting to demonstrate the positive impact of adaptive learning. Similar results have been observed through other research. Baker, Goldstein, and Heffernan (2011) studied middle school math students to determine the effect the intelligent tutoring program had on them. After observing 232 students over a year with over 580,000 transactions with the intelligent tutor, it was found that there was a strong correlation that indicated successful course results increased through the use of an intelligent tutor. Kennedy, Ioannou, Zhou, Bailey, and O'Leary (2012) studied medical students who completed simulated surgical trainings. In this study, the team created a prototype for an intelligent tutor that could be used to provide interactive feedback based on 48 variables within the following categories: general (timestamp, current stage); tool position (e.g., current force applied by drill); burr metrics; anatomical structure metrics; and bone specimen metrics. The team then used these course variables to develop a prototype of an ITS and associated dashboard. Mahmood, Hall, and Swanberg (2001) provides a meta-analysis of available ITS literature and found that, in the aggregate, ITS generally outperformed other modes of instruction other than individual and small group human tutoring.

Other research points to the potential application of adaptive training to support improved efficiency where performance gains remain constant. Long et al. (2015) developed an adaptive marksmanship training curriculum in which the training is modified individually based on previous performance. Interestingly, the research findings indicate that while overall performance was actually higher for the group receiving the standard training, the adaptive group completed the training in 40% less time overall with acceptable performance. Use cases such as these highlight the potential cost savings and efficiency of adaptive training systems. More generally, these evaluations indicate the need for a generalized framework for comparing individual offerings of a given course, such as marksmanship, and the effectiveness and efficiency of results across different courses and modalities.

Course Evaluation Framework

In order to develop the course evaluation tool, it was critical to determine how the BRM course could be evaluated. As Bramley and Newby (1984) write, there are five key purposes for evaluation: feedback (linking learning outcomes to objectives), control (linking training to organizational activities), research (determining the relationship between learning and transfer to the job), intervention (influencing the context), and power games (manipulating evaluative data for organizational decision making). The widely adopted Kirkpatrick model of evaluation (Kirkpatrick, 1994) was selected to encompass these evaluation considerations (see Figure 1). The model is based on four levels of evaluation of increasing scope. Level 1 and Level 2 evaluations were selected to demonstrate the methods and techniques for the purposes of this proof-of-concept while including the capability to add additional levels in the future.

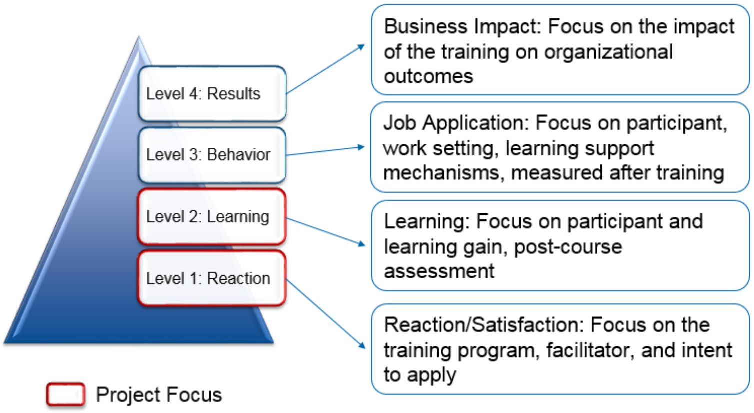

Level 4: Results
Level 3: Behavior
Level 2: Learning
Level 1: Reaction
Business Impact: Focus on the impact of the training on organizational outcomes
Job Application: Focus on participant, work setting, learning support mechanisms, measured after training
Learning: Focus on participant and learning gain, post-course assessment
Reaction/Satisfaction: Focus on the training program, facilitator, and intent to apply
Project Focus

Figure 1. Kirkpatrick Model for Evaluation

The Level 1 and 2 surveys and classroom tests were created to support the course evaluation tool and are not intended to represent curriculum recommended materials but rather illustrate tool capabilities. For Level 1, a reaction / satisfaction survey was adapted from an amalgamation of existing surveys developed for other client courses. The survey consisted of multiple items with a 7-point Likert-type scale and included three open-ended questions. For Level 2, which targets learner performance, ten multiple choice questions were developed based on the Army Field Manual on Rifle Marksmanship M16/M4-Series Weapons (FM 3-22.9). To provide the basis for a learning gain analysis, the research team assumed this test was given before and after the classroom portion of the BRM course for data simulation purposes. Additionally, the qualification rating based on record fire scores (DA Form 3595-R) was used as the performance measure for the live fire portion of the course. Qualification ratings were assessed based on a 40 round, multiple position evaluation with results noted as expert (36-40 hits), sharpshooter (30-35 hits), marksman (23-29 hits), and unqualified (less than 23 hits). Soldiers must maintain qualification to remain in the Army.

Literature was reviewed to determine the critical components of an effective dashboard. The components should be able to answer the questions that dashboard users have, and convey relevant information in a direct manner. Gasevic, Dawson, and Siemens (2015) describe utilization of course signals (like traffic lights) to indicate areas of concern. Corrin and de Barba (2014) found that the top impacts of a learning dashboard for students are: reflection on self-regulated learning; ability to plan new or amend strategies; self-motivation; peer-motivation (based on class average); and usefulness (being able to see everything in one place). Chatti, Dyckhoff, Schroeder, and Thüs (2012) developed a reference model for learning analytics that was used to orient the overall design and provides a tool for ensuring all perspectives are covered. Campbell and Oblinger (2007) describe the emerging field of academic analytics that includes relevant variables based on data for institutional-level analysis, i.e., above the individual course level. A number of additional references were utilized to identify variables of interest and approaches for dashboard displays, including Peterson (2016) and Chatti, Dugoija, Thüs, and Schroeder (2014).

Demographic Variables

Demographic variables are commonly used to explore and explain learner performance. Some demographic variables may be considered independent of training domain while others specific to a course. Education level is perceivably independent of the domain. In the case of the BRM course, prior rifle experience is perceivably unique to marksmanship. The literature on marksmanship was reviewed to identify variables that may explain learner performance that could be included in the proof-of-concept tool. Chung, Delacruz, de Vries, Bewley, and Baker (2006) identified five variable categories that affect rifle marksmanship performance: (1) perceptual-motor variables (steadiness, prior shooting experience, device-fire performance); (2) cognitive variables (training/instructional effects, aptitude, knowledge of shooting); (3) affective variables (confidence, anxiety, attitudes), and variables external to the learner; (4) equipment; and (5) environment. Within these categories, it can be seen that some variables are demographic in nature; i.e., describe a population. For example, prior shooting experience and knowledge of shooting are demographic variables. Lipinski, James, and Wampler (2014) found that prior knowledge did significantly predict performance albeit too small to group learners by prior knowledge. While anxiety is an affective variable, Chung, O'Neil, Delacruz, and Bewley (2005) also noted how anxiety can increase heart and breathing rates which in turn exacerbates position instability leading to poor performance. So while affective in nature, anxiety does have an effect on psycho-motor skills. Other researchers have looked at other possible explanatory variables such as hydration, diet, and caffeine, (Tharion & Moore, 1993; Tharion, Szlyk, & Rauch, 1989; Kamimoriet, Johnson, Belenky, McLellan, & Bell, 2004). Other marksmanship research was also reviewed from the perspective of what they perceived as explanatory demographic variables (Jenson & Woodson, 2012; Tierney, Cartner, & Thompson, 1979; Chung, Nagashima, Delacruz, Lee, Wainess, & Baker, 2011).

Based on this research, the demographic variables listed in Table 1 were selected to support the development of the course evaluation tool. These variables are not considered all-inclusive but rather representative of demographic variables that could be explored through the course evaluation tool. The variables are categorized as either standard, i.e., likely independent of domain and applicable to multiple courses, or as non-standard, i.e., specific to the BRM course. Again, the intention was to illustrate the tool's capability to accommodate common standard and course unique demographic variables. Note that the constructs of anxiety, self-efficacy, and motivation are considered standard; however, they are likely to be customized to the specific task performance. For example, it may be more informative to be aware of anxiety associated with shooting a weapon for live fire performance while anxiety associated with taking a course could be informative for classroom performance.

DATA GENERATION

A complete set of data did not exist for the selected use case, so the research team simulated the variety of source data required to populate the analysis. The simulated data fit within an overarching overall technology solution and can be generalized to other courses that use the xAPI or GIFT frameworks. This section summarizes the data, decisions made in simulating data, and the technical approach that accommodated the data plan.

Course data included several data sets: learner profiles; pre- and post-lesson knowledge assessment surveys; ability assessment (record fire); a student attitude survey; and a student reaction/satisfaction survey. Learner profile data could have been represented with an xAPI wrapper, but actually included any JSON (JavaScript Object Notation) description that uses a valid xAPI Agent identifier. Users were identified by an email address in the format learnerXX@example.com. Learner profiles represented statistically-relevant joint distribution factors such as 78:22 enlisted-to-officer ratio, education level, and age, following binomial (0 or 1) or skewed Gaussian distributions, as appropriate. Where available, references were used to identify population distributions, such as the publicly available Defense Manpower Data Center repository. Learner data can be extracted from a Learning Record Store (LRS), assuming they are stored as JSON files or are wrapped in valid xAPI statements.

For each learner profile generated, there was a corresponding simulation of the learner filling out pre- and post-lesson knowledge assessments. These assessment surveys were generated in GIFT using GIFT's survey authoring tool, then exported from GIFT using the native utility. The simulation read the exported GIFT survey (exported in JSON format), created a data frame to hold student responses, and then simulated responses using input parameters from the learner profile, such as a student's own assessment of his firearm experience (rifle and pistol). A combination of maximum likelihood estimation and Kolmogorov kernel density estimation was used to predict correct responses based on mean and variance statistics from reference data. This process was repeated for post-lesson knowledge assessment with increased likelihood of correct answers. The simulation then scored each assessment, using the scoring rubric defined in the GIFT survey, and calculated the improvement for each learner. In a real-world application, these assessments might be stored in xAPI format in an LRS. For the sake of simplicity, we stored a JSON object for each assessment for each learner directly to disk or to our analysis database.

The record fire qualification data simulation followed a similar procedure. For each learner, the simulation predicted a number of "hit," "miss," and "no fire" outcomes based on students' previous rifle experience. A standard 40 round sequence in three firing positions was simulated at various ranges. The simulation also included approximately 5% of students not qualifying in the record fire evaluation exercise and reattempting at another date. The qualification rating the students earned in the exercise was a joint distribution based on prior experience and amount of practice time. Accuracy (shot groups) was not simulated to correspond to marksmanship ratings; only the number of hits. The results of the simulated record fire exercise were written into individual JSON files to represent the way data would be stored from Range Experience Acquisition Portal for Evaluation & Reporting (REAPER) in the xAPI format (Durlach, Washburn, & Reagan 2015). For convenience, the complete files were also written to disk in Comma-Separated Values (CSV) format. Data files were again stored in the analysis database.

To simulate students' anxiety, motivation, and self-efficacy, responses were simulated for these sections of the GIFT survey based on a learner's education level. As education level increases, a student's answer to questions in all three categories decreases along a Likert range (in this case, a discrete score 1-7). A distribution was generated for each variable with its own mean and variance, and then repeatedly sampled to answer questions for each student. Again, results were stored in JSON and CSV formats to represent both the real-world process and simplified data management for convenience.

The analysis framework is independent of the data simulation but drove many of the data simulation decisions. For each analysis, our goal was for the analysis tools to be agnostic of the course and course surveys. In most cases, it relied on self-describing data formats, such as JSON and xAPI. This goal was achieved in most, but not all, analyses. There was only one use case for the proof-of-concept, but a future version will allow users to identify all of their data sources (with xAPI end-points or by uploading data) and the system will organize the data for appropriate analyses. In all cases, data was collected into a database to pre-process the data and recommend features for evaluation. Interim results were stored into associated JSON files so that users can extract information for further research, if desired. This approach had the added benefit of improving website response times and enables interactive data visualizations.

APPROACH AND RESULTS

This section describes the results of incorporating the data simulated for this research within the analytic system. The approach is referred to as the Health Indicators area of the dashboard as it provides a high-level summative view of the course performance measures that allow for issues to be identified rapidly. The specific statistical procedures utilized to produce the results shown here are described below, although the system will ultimately allow the advanced user to select and configure their own analytic methods. The methods were selected based on their accessibility and ease of use for demonstrating the applicability of the concepts herein and specifically for the data sets generated as described above. Future iterations will focus on development of an automated pipeline that adapts instruction to the observed data and selects the best methods to populate the Health Indicators while not allowing the user to apply methods that violate critical assumptions or are not otherwise appropriate.

The user can readily drill down to understand performance according to these performance measures based on descriptive and inferential statistics. The Health Indicators area is intended to intelligently and efficiently support instructors in course evaluation based on learner performance with minimal direct references to statistical outputs unless desired by the user. The course performance measures in the Health Indicators are created by the user by selecting from a list of potential measures and student data when the course is initially created. The colors indicate potential issues in a stop-light color format and would be scaled according to the instruments used. As shown in Figure 2 for the Basic Marksmanship course, the selected health indicators were Attitude, Reaction/Satisfaction, Post-Test Scores, and Qualified. Attitude includes motivation, self-efficacy, and anxiety and is populated in these categories directly from GIFT based on a Likert-type student survey. In the proof-of-concept, the survey was administered within GIFT, although external data would work equally as well.

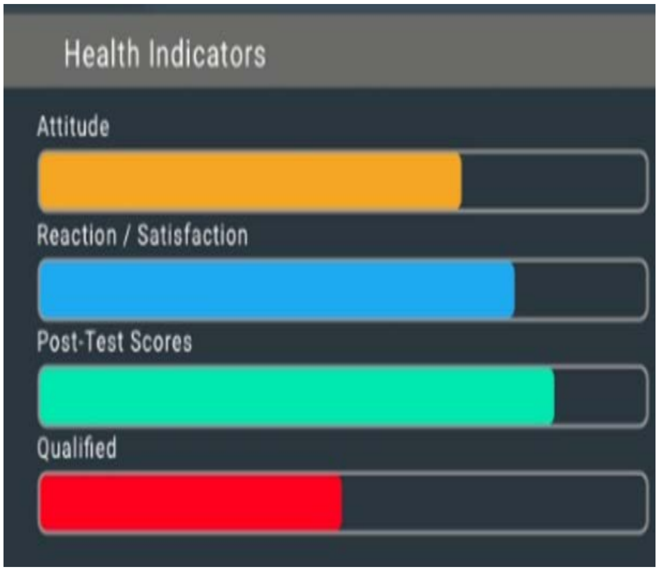


Figure 2. Health Indicators area for Basic Marksmanship course

Reaction/Satisfaction is collected through another survey administered directly through GIFT. Reaction/Satisfaction is a Kirkpatrick Level 1 course evaluation measure intended to collect student perceptions of the course, commonly comprised of Likert-type survey items with some open-ended questions. Post-Test Scores is a common learner performance measure, and in this instance, the average post-test score is provided. Other courses, or other instructors, may prefer to see other performance measures as a health indicator such as average change in scores. Again, this could be selected when the course was initially created within the framework. The last health indicator, Qualified, is a performance measure unique to a marksmanship course that indicates the percentage of learners who qualified based on Record Fire performance measure; i.e., got at least 23 out of 40 target hits. While seemingly unique to the Basic Marksmanship course, this type of binary variable (go/no go; pass/fail) could be applicable to other certification type courses.

Each health indicator is an interactive link that provides the user with more information about the selected indicator. In the case of Post-Test Scores, the user would be presented with the average pre-test, average post-test, and average change in scores as shown in Figure 3. For this data set, a t-test is utilized to determine if two data sets exhibit a statistically significant difference at the 95% level of confidence. Note that level of confidence and other details are made available to users that select the “How to Interpret” box on the screen but is otherwise not directly referenced to focus on the actionable results.

In addition to the average scores, the user is also presented with a table that indicates which demographic variables have different sample means, if any, as shown in Figure 4. The table is generated by iteratively applying analysis of variance (ANOVA) using the selected score change as the dependent variable and each demographic variable as the independent variables. Tukey's post hoc analysis is then applied to independent variables that have statistically different means to rank order them based on omega squared (ω^2) values. While other statistical tests could ultimately be more definitive by accounting for covariates and partial correlations, for purposes of this proof of concept, multiple one-way ANOVA was utilized for simplicity. Independent variables with statistically significant contributions to the variance, p < .05, are binned based on large contribution (ω^2 => .14) or moderate contribution (.06 <= ω^2 < .14). This data table provides an interactive mechanism for the user to explore variables that might help explain learner performance. Instead of exploring all demographic variables to assess affects, the user is directed to those that had a statistically significant large or moderate contribution to the explanation of the performance measure, in this case test scores. Intuitively, this indicates that the user should explore these factors to determine what explains performance on test scores.

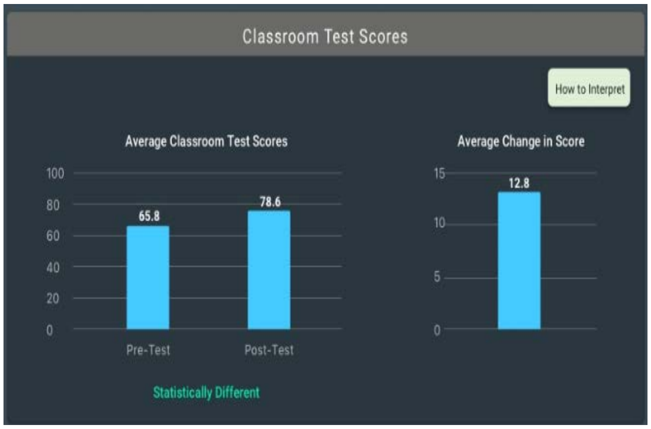
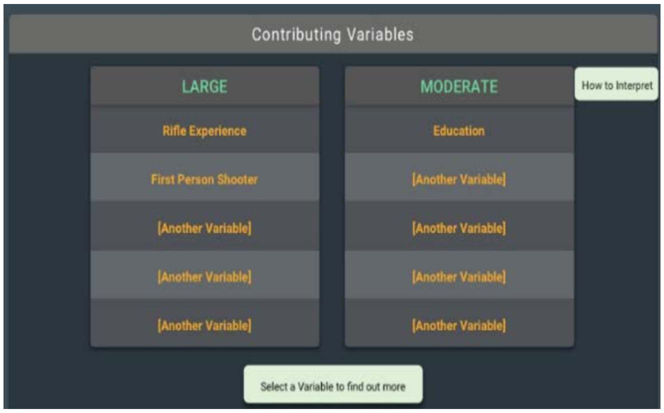


Figure 3. Classroom Test Scores Shown after Selecting Post-Test Scores from Health Indicators area


Contributing Variables

How to Interpret

LARGE	MODERATE
Rifle Experience	Education
First Person Shooter	[Another Variable]
[Another Variable]	[Another Variable]
[Another Variable]	[Another Variable]
[Another Variable]	[Another Variable]

Select a Variable to find out more

Figure 4. Demographic Variables with a Statistically Significant Contribution to the Explanation of a Performance Measure

The variables in the table are interactive links to further investigate that contributing variable by exploring the groups within the variable. For example, as shown in Figure 5, after selecting the demographic variable, Rifle Experience, five groups ranging from Never Fired to Expert are displayed, with the average change in score for each group. The user can see that those who Never Fired a rifle have the greatest positive change in score while those with Sharpshooter exhibited no change in score and Expert actually incurred a negative change in score. So while the average change in score was positive for the course, exploring the groups reveals that some groups experienced more positive change than other groups. Furthermore, ANOVA is applied again to inform the user whether differences between groups are statistically significant at the 95% level of confidence, as shown in Figure 6 for the Novice and Marksman groups. This type of statistical analysis and data presentation helps the instructor better understand performance and take action to provide remedial help or make course improvements to support low performing groups. This type of analysis helps to answer important questions from an instructional strategy, such as: does student performance improve; do all groups improve similarly or do some improve more than others; and, crucially, what might help explain these differences in performance?

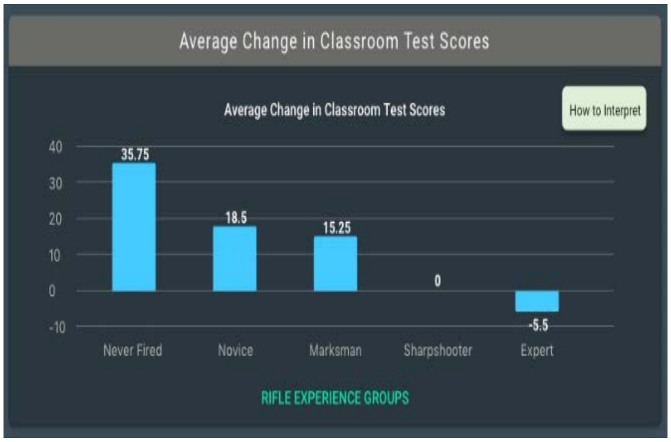
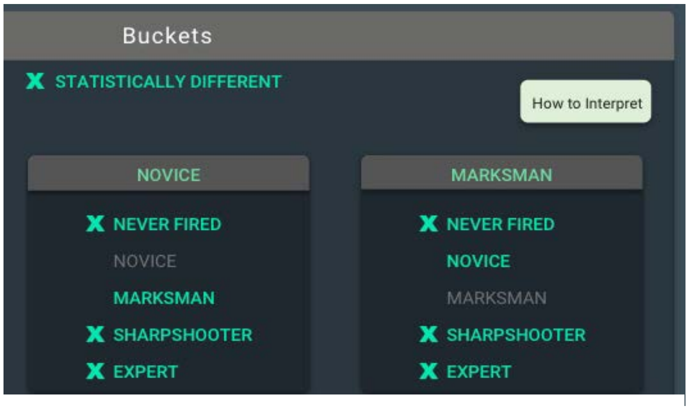


Figure 5. Average Change in Scores by Group for the Demographic Variable Rifle Experience

The other performance measures in the Health Indicators area also have statistically-driven drilldown capabilities, but depending on the variable type, different procedures are applied. In the previous case of test scores, this dependent variable is continuous in nature and therefore lends itself to using ANOVA when exploring categorical demographic variables as the independent variable. In the case of Qualified, the dependent variable is a categorical Qualification Rating of Expert, Sharpshooter, Marksman, or Unqualified. In this case, a Chi-Square for Association with post-hoc multiple comparison via adjusted standardized residuals is used to determine where statistically significant differences occur between observed values and expected values. The expected values are the number of observations one would mathematically expect if there were no association between the categorical dependent and categorical independent variables. This statistical test is performed between the Qualification Rating and all categorical demographic variables. Those demographic variables that have a statistically significant relationship with Qualification Rating are listed in table format as shown in Figure 6. In our example, Rifle Experience was selected to further explore the relationship between Qualification Rating and the groups within Rifle Experience. As indicated in Figure 7, the observed and expected number of students are presented by Qualification Rating for each of the Rifle Experience groups.



Buckets

✗ STATISTICALLY DIFFERENT

How to Interpret

NOVICE	MARKSMAN
✗ NEVER FIRED	✗ NEVER FIRED
NOVICE	NOVICE
MARKSMAN	MARKSMAN
✗ SHARPSHOOTER	✗ SHARPSHOOTER
✗ EXPERT	✗ EXPERT

Figure 6. Statistically Different Groupings with Respect to Marksmanship Performance

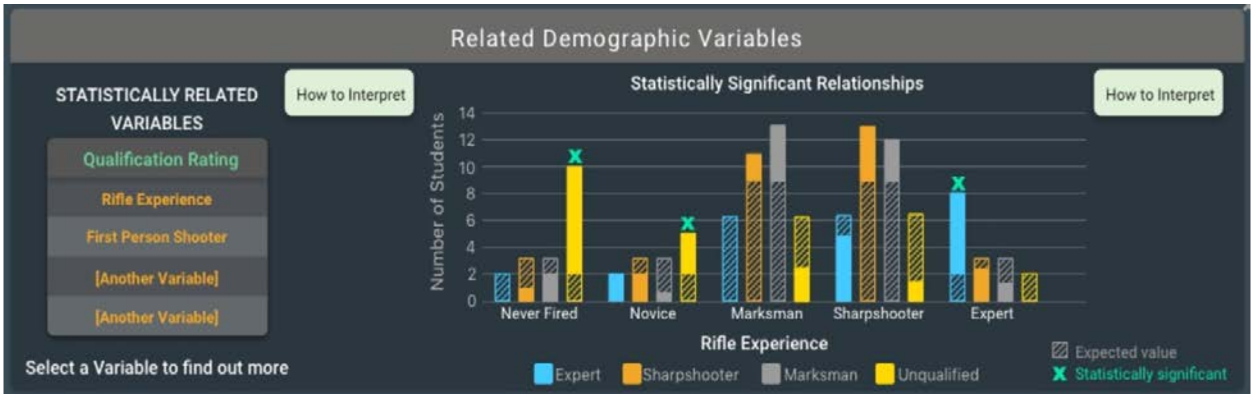


Figure 7. Chi-square for association between Qualification Rating and the demographic variable Rifle Experience

In this example, it can be interpreted that those who “Never Fired” for Rifle Experience are associated with those who have a Qualification Rating of Unqualified. In addition to providing a data visualization with intrinsic value, applying this procedure helps to answer the question of whether expected relationships play out in the observed results. In this case, students entering the course in the Never Fired experience level are disproportionately rated as Unqualified, indicating that this demographic is an area to explore further from an instructional strategy perspective. For these perceivably complex data interpretations, the use of the How to Interpret interaction becomes important for those not familiar with statistical inferences. For the Chi-square for Association statistical test, the How to Interpret includes interpretative statements that are automatically generated based on the results to help guide the user. In this case, the generated statements would read:

- Never Fired, Novice Rifle Experience students are more likely to be Unqualified.
- Expert Rifle Experience students are more likely to be Expert.

In summary, the prototype currently supports descriptive statistics and inferential statistics (t-tests, one-way ANOVA with post-hoc Tukey, and Chi-square for Association with post-hoc comparisons) to explore relationships between dependent and independent variables that are continuous or categorical in nature. This structure forms the basis for an automated analysis pipeline that will eventually select the most appropriate test based on fit with observed data, once actual data in GIFT becomes available. In each case, the dashboard tries to display useful data regardless of the presence of statistically significant relationships to ensure that the displays will be of value even for small data sets or those without a statistically quantifiable relationships. It must be kept in mind that these procedures provide guidance in understanding performance but do not indicate causation.

The statistical engine was developed with the flexibility to add other methods and tests as new research questions and methods are identified. Current data presentations include horizontal and vertical bar charts, 100% filled vertical bar charts, and scatter plots. Other data presentations were considered such as pie charts but from a user experience perspective, bar graphs are recognized as easier to interpret results. The interface was specifically designed from a calming color palette within a clean, simplistic interactive experience that is easy to use and affords multiple means of navigation.

FUTURE RESEARCH

The tools and technology necessary to create rich views of student performance are becoming more widely available and accessible every day. While the components of effectiveness evaluation will likely remain fairly consistent over time, the specific methods and techniques for evaluation will continue to evolve. By building an open source architecture that utilizes tools with active development communities, future applications can be built to address use cases as they arise. As described in Long et al. (2016), a myriad of opportunities for data analytics applications exist to be experimented with and incorporated into the tool suite. Examples include the use of predictive modeling to enhance identification of at-risk students that incorporates prior performance data, social network analysis that draws from social media, sentiment analysis of user feedback, and a host of other examples. As evident from the approach above, extension of the analysis to address behavior change and performance interventions could be a natural next step.

An important limitation is the availability of data in a format conducive to being imported and processed with minimal effort. While xAPI provides a framework for interoperability, Army learning data resides on a host of systems with different specifications and the Sharable Content Object Reference Model (SCORM) remains widely utilized in learning systems. Commercially developed learning management systems can also utilize proprietary data formats that are not conducive to data transfer. One approach to wider adoption of xAPI as a standard, as modeled by the Advanced Distributed Learning (ADL) Initiative, is open and wide-spread publication of effective tools that make use of these data analysis capabilities. If tool sets are shared freely on platforms that are widely used in the data analysis community, e.g., GitHub, then the relatively lower cost of use will lead to broader adoption. Alongside this approach must be a demonstration of the potential and capabilities that utilize diverse and interoperable data sets made possible by standards such as xAPI. This process has led to wide-spread commercial adoption of the group of methods coined predictive analytics, which over time have made the integration of disparate data sources a source of competitive advantage to such a degree that the competitive environment in many sectors requires an analytics capability (Davenport, 2006; Rivera, 2015). If the Training and Education community progressively sets the standard for effectiveness evaluation by utilizing interoperable, diverse data sets and demonstrates the efficacy of these techniques, then over time, expectations will lead to the broader option of standards such as xAPI.

ACKNOWLEDGEMENTS

The researchers would like to acknowledge the following individuals: Paul Cummings for providing the vision and guidance to develop an intuitive user experience; Gary Haus for developing the artwork and design that underpin the models described above; Raghav Joshi for creating the initial data generation pipeline; Robbie Brusso for his vision for the statistical engine components; and finally, Janet Weisenford, Jan Cannon-Bowers, and Kristy Murray for their review and feedback on the research approach.

REFERENCES

Baker, R. S. J. D., Goldstein, A., Heffernan, N.T., (2011). Detecting Learning Moment-by-Moment. *International Journal of Artificial Intelligence in Education, vol. 21, no. 1-2,* 5-25. Retrieved from http://www.columbia.edu/~rsb2162/BGH-IJAIED-v29.pdf

Bollin, A. (1998). Improving Lectures and Practical Classes in Using an Automatically Feedback System. WebNet 98 World Conference of the WWW, Internet, and Intranet Proceedings.

Bramely, P. & Newby, A.C. (1984). The Evaluation of Training Part I: Clarifying the Concept. *Journal of European & Industrial Training, 8,6,* 10-16.

Campbell, J.P. & Oblinger, D.G., (2007), Academic Analytics. *Educause*. October 2007.

Chatti, M.A, Dugoija, D., Schroeder, U., & Thüs, H. Learner Modeling in Academic Networks. (2014) IEEE 14th International Conference on Advanced Learning Technologies, pp. 117-121.

Chatti, M.A, Dyckhoff, A.L., Schroeder, U., & Thüs, H. A (2012). Reference Model for Learning Analytics. *International Journal of Technology Enhanced Learning 4(5/6)*, pp. 318-331.

Chung, G. K. W. K., Delacruz, G. C., de Vries, L. F., Bewley, W. L., & Baker, E. L. (2006). New directions in rifle marksmanship research. *Military Psychology, 18,* 161-179.

Chung, G. K. W. K., Nagashima, S. O., Delacruz, G. C., Lee, J. L., Wainess, R., & Baker, E. L. (2011). Review of rifle marksmanship training research. The National Center for Research on Evaluation, Standards, and Student Testing, CRESST Report 783, University of California, Los Angeles.

Chung, G. K. W. K., O'Neil, H. F., Jr., Delacruz, G. C., & Bewley, W. L. (2005). The Role of Affect on Novices' Rifle Marksmanship Performance. *Educational Assessment, 10,* 257-275.

Corrin, L., & de Barba, P. (2014). Exploring Students' Interpretation of Feedback Delivered Through Learning Analytics Dashboards. In B. Hegarty, J. McDonald, & S.-K. Loke (Eds.), *Rhetoric and Reality: Critical perspectives on educational technology.* Proceedings of Ascilite Dunedin Conference 2014, 629-633. Retrieved from http://ascilite2014.otago.ac.nz/files/concisepapers/223-Corrin.pdf

Davenport, T. (2006). Competing on Analytics. *Harvard Business Review*, retrieved on 2 June 2016. https://hbr.org/2006/01/competing-on-analytics

Durlach, P., Washburn, N., & Regan, D. (2015). Putting Live Firing Range Data to Work Using the xAPI. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), 2015, Paper No. 15019.

Foxon, M. (1989). Evaluation of training and development programs: A review of literature. *Australian Journal of Educational Technology, 5(2),* 89-104.

Gasevic, D., Dawson, S., & Siemens, G. (2015) Let's not forget: Learning analytics are about learning. *TechTrends, 59 (1),* 64-71.

Hruska. M, Amburn, C., Long, R., Kilcullen, T., Poeppelman, T.R., (2014). Experience API and Team Evaluation: Evolving Interoperable Performance Assessment. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), 2014, Paper No. 14157.

Jensen, T., & Woodson, J. (2012). A naval marksmanship training transfer study - The use of indoor simulated marksmanship training to train for live fire. (Master's Thesis). Naval Post-Graduate School. Monterey, CA.

Kamimori, G. H., Johnson, D., Belenky, G., McLellan, T., & Bell, D. (2004). Caffeinated gum maintains vigilance, marksmanship, and PVT performance during a 55 hour field trial. Walter Reed Army Institute of Research, Washington, D.C.; Defence R&D of Canada, Toronto, Ontario.

Kennedy, G.E., Ioannou, I., Zhou, Y., Bailey, J. & O'Leary, S. (2012) Data mining interactions in a 3D immersive environment for real-time feedback during simulated surgery. In M. Brown, M. Hartnett & T. Steward (Eds.), Future challenges, sustainable futures. Proceedings of Ascilite Wellington Conference 2012, 468-478. Retrieved from http://www.ascilite.org/conferences/Wellington12/2012/images/custom/kennedy,_gregor_-_data_mining.pdf

Kirkpatrick, D. L. (1994). *Evaluating training programs: The four levels*. San Francisco: Berrett-Koehler.

Lipinski, J. J., James, D. R., & Wampler, R. L. (2014). Test of prior marksmanship knowledge predictor test. United States Army Research Institute for the Behavioral and Social Sciences, Research Report 1970.

Long, R., Hruska, M., Medford, A., Murphy, J., Newton, C., Kilcullen, T., Harvey, R.L. (2015), Adapting Gunnery Training Using the Experience API. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), 2015, Paper No. 15179.

Long, R., Smith, M., Dass, S., Dillon, C., & Hill, K. Data Analytics: Techniques and Applications to Transform Army Learning. (2016). ModSim World Conference, NDIA, 2016.

Mahmood, M.A, Hall, L., & Swanberg, D.L. (2001) Factors Affecting Information Technology Usage: A Meta-Analysis of the Empirical Literature, *Journal of Organizational Computing and Electronic Commerce, 11(2).*

Peterson, J.L, (2016), Formative Evaluations in Online Classes. *The Journal of Educators Online, ISSN 1547-500X, Vol 13, Number 1.*

Rivera, J. (2015). Gartner Says Retailers Need Advanced Analytic Capabilities to Compete in the Digitalized Marketplace. Gartner, retrieved from: http://www.gartner.com/newsroom/id/3093819

Tharion, W. J., & Moore, R. J. (1993). Effects of carbohydrate intake and load bearing exercise on rifle marksmanship performance. United States Army Research Institute of Environmental Medicine, Technical Report No. TR-5-93. Natick, MA.

Tharion, W. J., Szlyk, P. C., & Rauch, T. M. (1989). Fluid loss and body rehydration effects on marksmanship performance. United States Army Research Institute of Environmental Medicine, Report No. M57-89. Natick, MA.

Tierney, T. J., Cartner, J. A., & Thompson, T. J. (1979). Basic rifle marksmanship test: Trainee pretest and posttest attitudes. United States Army Research Institute for the Behavioral and Social Sciences, Technical Paper 354, Alexandria, VA.