ARMY RESEARCH LABORATORY

ARL

An Evaluation of the Generalized Intelligent Framework for Tutoring (GIFT) from a Researcher's or Analyst's Perspective

by Robert A Sottilare, Anne M Sinatra, Julian Watson, Zachary Davis, Stefani King, and Michael D Matthews

ARL-SR-0302   **December 2014**

NOTICES

Disclaimers

Army Research Laboratory

Aberdeen Proving Ground, MD 21005-5425

An Evaluation of the Generalized Intelligent Framework for Tutoring (GIFT) from a Researcher's or Analyst's Perspective

Robert A Sottilare and Anne M Sinatra
Human Research and Engineering Directorate, ARL

Julian Watson, Zachary Davis, Stefani King, and Michael D Matthews
United States Military Academy

REPORT DOCUMENTATION PAGE			*Form Approved* *OMB No. 0704-0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**			

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*	
December 2014	Final	20 January–4 February 2014	

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
An Evaluation of the Generalized Intelligent Framework for Tutoring (GIFT) from a Researcher's or Analyst's Perspective	
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Robert A Sottilare, Anne M Sinatra, Julian Watson, Zachary Davis, Stefani King, and Michael D Matthews	
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
US Army Research Laboratory ATTN: RDRL-HRT-T Orlando, FL 32826	ARL-SR-0302

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Current US Army standards for training and education are group instruction and classroom training, also known as one-to-many instruction. Recently, the Army has placed significant emphasis on self-regulated learning methods to augment institutional training. Per the Army Learning Model, Soldiers will be largely responsible for their own learning. One-to-one human tutoring has been shown to be significantly more effective than one-to-many instruction but it is not practical to assign each Soldier a personal mentor. An alternative to one-to-one human tutoring is one-to-one computer-moderated tutoring using artificially Intelligent Tutoring Systems (ITSs), which have been shown to be effective in promoting individual learning in static, simple, well-defined domains (e.g., mathematics). To be practical, high authoring costs and limited adaptiveness barriers must be addressed. This report describes the outcomes of an evaluation conducted at the US Military Academy to determine initial usability of the Generalized Intelligent Framework for Tutoring, a tutoring architecture constructed with the goal to reduce time and skill needed to construct ITSs while increasing their adaptiveness or ability to act autonomously to optimize user learning. Participating cadets were assigned tasks related to the researcher's or analyst's perspective as part of a course assignment in PL488E, an engineering colloquium. Their thoughts are shared herein along with technical challenges identified by the US Army Research Laboratory based on cadet observations.

15. SUBJECT TERMS

self-regulated learning methods, intelligent tutoring system, Army Learning Model, adaptive tutoring, effect evaluation, usability

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Robert A Sottilare
a. REPORT	b. ABSTRACT	c. THIS PAGE	UU	16	19b. TELEPHONE NUMBER *(Include area code)* (407) 208-3007
Unclassified	Unclassified	Unclassified			

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

INTENTIONALLY LEFT BLANK.

1. Introduction

The standard for acquiring knowledge in institutional training within the US Army is split between traditional classroom training and live training. These methods are used to test recall and allow Soldiers to apply and test their skills respectively in varying conditions and against a set of standards. Over the last 30–40 years, virtual simulation has been added to the training toolbox and a debate has raged about what mix of live and virtual training is optimal. To augment institutional training, and provide flexibility and accessibility for Soldiers who need training, the Army has recently emphasized self-regulated learning; Soldiers are largely responsible for managing their own learning. From a common sense point of view, it may not seem practical for each Soldier to be able to manage his/her learning without some guidance. This guidance, also referred to as coaching, mentoring, or tutoring, is usually provided one to one by a human tutor. Generally, this function has fallen upon noncommissioned officers. However, the success of one-to-one tutoring recognized by Bloom (1984, 2σ effect size) and VanLehn (2011, 0.8σ effect size) are impractical to implement in large organizations like the Army.

Once we decide to pull the human tutor out of the instructional loop, our alternative is to provide one-to-one computer-guided instruction using intelligent tutoring systems (ITSs), which have been shown to be effective in promoting individual learning in static (e.g., desktop), simple, well-defined (procedural) domains (e.g., mathematics, physics). Well-defined domains generally have one solution to a problem presented whereas ill-defined domains may have multiple paths to success. ITSs are a practical alternative to one-to-one tutoring but are costly to author (develop) and do not have sufficient adaptability to support more dynamic, complex, ill-defined domains represented in many Army operations. To address the needs of learners, authors, and analysts/researchers who use or might use adaptive tutoring technologies to learn, develop new ITSs, and analyze the effect of ITS technologies, the US Army Research Laboratory (ARL) created the Generalized Intelligent Framework for Tutoring (GIFT) (Sottilare et al. 2012).

GIFT is a prototype open-source, service-oriented, adaptive tutoring architecture targeted to support automated authoring, automated one-to-one and one-to-many guided instructional experiences, and evaluation of effect to determine the impact of current and emerging tutoring technologies with regards to learning outcomes. Ultimately GIFT will be a community development project. Currently there are about 400 users in 30 countries who are registered users of GIFT, which is freely available at www.GIFTtutoring.org.

This report is one of 3 evaluating the usability of GIFT from 3 perspectives: learners, authors, and researchers/analysts. This report is focused on the researcher's/analyst's perspective, which is about what people who use GIFT to evaluate adaptive tutoring technologies (e.g., tools and

methods) think about their experience and GIFT's ease of use in facilitating and managing their experiment/evaluation planning, execution, and data analysis.

The evaluation construct of GIFT is intended to provide tools and methods in a testbed environment to make it possible to easily evaluate the effect (e.g., learning effect, performance effect) of various ITS technologies and methods. GIFT currently supports experimental design, data collection, and evaluation as a testbed function to compare/contrast the effectiveness of adaptive tutoring technologies.

This report outlines researcher/analyst evaluations conducted by cadets within the Engineering Psychology Program, part of the Behavioral Science and Leadership Department at the US Military Academy (USMA) as part of their coursework in "Human Factors of Military Training Simulations" (PL488E) during the 2014 Spring semester.

2. Evaluation of GIFT from a Researcher's or Analyst's Perspective

2.1 Introduction

Self-efficacy scores and mood differences were measured as an outcome of taking the GIFT logic puzzle. They were effective in viewing any changes to initial baseline ratings. GIFT, being a completely new and innovative concept, was being tested to view the self-efficacy and mood differences at various times to help gauge future testing procedures in an attempt to raise both scores. Though self-efficacy scores and mood seemed promising, with the exception of one time measurement, they showed almost no statistical significance. They were relatively volatile and hard to analyze because of the rather vague interpretation of mood and self-efficacy coupled with the low sampling rate of 3 participants (users).

2.2 Research and Analysis Using GIFT

In a growing digital age, researchers are constantly looking for new ways to learn tasks in an effective and low-cost manner. In the Army, where a diminishing budget is a real threat to efficient training, researchers are increasingly searching for innovative ways to minimize cost while maximizing Soldier preparation for war. GIFT, an ITS, shows promise as a potential solution for effective learning outside of the traditional Army classrooms and encompasses 4 major components that define the way in which it actively teaches and reiterates knowledge to a new user (Pavlik et al. 2013).

The first component is known as the domain model, as in a distinguishable set of skills and knowledge and "is a representation of all the possible student states in the domain" (Pavlik et al. 2013, p. 39). The second component is the student model. It is distinguished as the subset of the domain model and one that can change throughout the course of learning. In other words, human states throughout the testing can be inferred and interpreted though provided performance

data. The third component is effectively known as the pedagogical model, which is interesting in the sense that it is not a static model; in fact, it is a model meant to be fluid and changing based on the needs of the user at any specific point in time of the intelligent tutoring session. The fourth and final component is known as the tutor-student interface model and is unique because of its changing media output that is based on the media input of the user. GIFT uses and employs all 4 models with the addition of an optional sensor module in an attempt to optimize learning potential (Pavlik et al. 2013). In the case of the logic puzzle, GIFT used a variety of the components to help the participant actively engage and solve the problem.

Moreover, understanding GIFT and the implications of human-less tutoring systems may actually give insight into the way in which humans perceive the concept of artificial intelligence in the future (Sottilare 2013). By using GIFT as a stepping stone, the future of teaching has the potential to be completely overhauled. Statistics derived from mood ratings and self-efficacy scores give insight into the user's experience and create future perceptions. In summary, as stated by Sottilare, "theoretical concepts of today will evolve into the practical implementations of tomorrow" (Sottilare 2013, p. 195). Furthermore, when developing ITSs, analyses must be performed on empirical data to develop a variety of key functions to include measures of success and adaptive instruction and support (Sottilare 2013).

2.3 Methods

The following sections describe the participants, apparatus, procedure, and results of the evaluation of GIFT as a research and analysis tool.

2.3.1 Participants

The evaluators of GIFT authoring tool usability are "firsties" (senior-level cadets) at USMA.

2.3.2 Apparatus

The apparatus used in this evaluation was the intelligent tutoring system architecture, GIFT, and specifically the Event Reporting Tool (ERT), to collect user data.

2.3.3 Procedure

First, 3 participants ("users", 2 male and 1 female) between the ages of 21 and 29 (mean [M] = 24, standard deviation [SD] = 4.36) ran through the GIFT Logic Puzzle Tutorial on an Alienware M17x laptop. Each participant created an individual profile and completed the tutorial in a single session, each session taking about 1 hr. The sessions consisted of pre-surveys, the tutorial, mid-surveys, assessment questions, a logic puzzle, and post-surveys. The tutorial and logic puzzle consisted of matching various food items to their purchasers based on different types of clues given in the game. A practice test was given between the 2 scenarios focusing specifically on familiarizing the user with the various clues.

The surveys gauged mood and self-efficacy, the latter measured using a Self-Efficacy Questionnaire (SEQ). The surveys were administered before the tutorial (time 1), halfway through the tutorial (time 2), and after the logic puzzle had been completed (time 3). The data generated by the sessions was then extracted into a Microsoft Excel spreadsheet using the ERT. Each user's data was organized into rows and the measured variables (mood/SEQ at times 1, 2, 3, and test score) divided. The appropriate edits, such as summing the SEQ scores and grading the tests, were then made in Excel. Once these values were found, they were evaluated for their means and standard deviations. Then the moods and SEQs were tested for statistical significance using a paired-samples t-test. We ran these tests in both Excel and IBM's Statistical Package for the Social Sciences (SPSS) software. The spreadsheet output was then transferred to a .pdf file and the results were written up and analyzed.

2.3.4 Results

The statistical tests for mood supported the null hypothesis, which predicted that the values for the measured quantity would not change significantly during the logic puzzle course. According to a paired-samples t-test, there was no significant difference found between Mood 1 ($M = 35.3$, $SD = 26.7$) and Mood 2 ($M = 32.7$, $SD = 23.2$), $t(2) = 0.01$, $p = 0.845$. According to a paired-samples t-test, there was no significant difference found between Mood 2 ($M = 32.7$, $SD = 23.2$) and Mood 3 ($M = 23.0$, $SD = 24.6$), $t(2) = 0.01$, $p = 0.531$. Finally, according to a paired-samples t-test, there was no significant difference found between Mood 3 ($M = 23.0$, $SD = 24.6$) and Mood 1 ($M = 35.3$, $SD = 26.7$), $t(2) = 0.01$, $p = 0.651$.

The SEQ showed similar results to mood except for in one case. It was hypothesized that self-efficacy would increase significantly over time during the logic puzzle course. As participants continued through the course, they should have grown more confident in their abilities. However, some of the data did not reflect the hypothesis. According to a paired-samples t-test, there was no significant difference found between SEQ 1 ($M = 33.0$, $SD = 1.0$) and SEQ 2 ($M = 42.0$, $SD = 9.54$), $t(2) = 0.01$, $p = 0.244$. According to a paired-samples t-test, there was no significant difference found between SEQ 2 ($M = 42.0$, $SD = 9.54$) and SEQ 3 ($M = 46.7$, $SD = 2.52$), $t(2) = 0.01$, $p = 0.565$. According to a paired-samples t-test, there was actually a significant difference in the expected direction found between SEQ 3 ($M = 46.7$, $SD = 2.52$) and SEQ 1 ($M = 33.0$, $SD = 1.0$), $t(2) = 0.01$, $p = 0.009$. In all except for one case the general trend was SEQ increased as time increased.

3. Conclusions

Both variables behaved in interesting ways. Mood behaved exactly as expected, in that it supported the null hypothesis that there were no significant differences. SEQ, however, behaved unexpectedly. It supported the null hypothesis between times 1 and 2 and times 2 and 3 but rejected the null hypothesis and showed significance between times 1 and 3. After statistical analysis it was confirmed that mood ratings were inconsistent. This makes sense because mood is a volatile, nebulous, and subjective quality that is hard to quantify. There was also no specific mood the test looked to invoke. Some people may have been happy at the beginning of the test while others may have been upset. The same could be said of how they finished, with some finishing happier or more upset than others. One would expect SEQ values to have had statistical significance because it makes sense that self-efficacy increases as time increases. However, that was not the case between SEQ 1 and 2 and between SEQ 2 and 3. This is because one of the user's SEQ score actually decreased between the second and third times. The test between SEQ 1 and 3 showed the expected rejection of the null hypothesis and confirmed that the values were significant; self-efficacy increased as a direct effect of experience. If the sample was large, the results for mood would likely also show no significant differences. The results for SEQ would have more than likely shown more rejections of the null hypothesis. The content questions test had a perfect score of 20 points, and overall the participants did well. There was a mean score of 17.0 (85%) with a standard deviation of 1.0. The tutorial did appear to be both relevant and helpful, and there was only a single outlier, which was user 1's SEQ for time 3.

Many interesting observations were made pertaining to GIFT, a program our group used for the first time. On the whole, the actual test and survey taking, once in the GIFT program, was relatively easy and intuitive. However, getting to the tests or even opening the correct files proved difficult and cumbersome. There were a few organizational issues regarding the interface and design in general. Other than generally being clunky, no specific icons were used to access the program; instead, a series of folder operations and pathways were used to access a systematic approach to opening the launch screen. Assuming the polished GIFT product does not include sifting through umpteen levels of confusing files, the only real prerequisite skills needed for GIFT are a general familiarity with basic computer controls. In other words, inexperienced users would need nothing more than a help guide to navigate and understand the complexities of the system itself.

It was difficult and confusing to get the correct data from the ERT, and the data were presented in a bizarre fashion. To begin with, the ERT was not intuitive to set up and extract data from. To get the data we needed, we had to select options that were not selected and unselect options that were selected. While this process may not have been difficult to someone who knows the system,

it will certainly be an issue for new users. Recognizing that some of these options were most likely for more advanced functions and operation, the ERT could be improved by inserting buttons that allow simple functions such as t-tests to be completed with a single click.

Additionally, the output produced in Microsoft Excel was extremely problematic and produced doubles in every facet, whether question or numeric answer. This output was also set up with 3 columns and a large number of rows. This made scrolling though the data difficult. Some improvements might be to make it easier to select the data you desire, remove the duplicates, and show the output in a system with more rows than columns. It would also help if Microsoft Excel or IBM SPSS functions could be directly used by the program, much like Microsoft PowerPoint is used during Logic Tutorial. Thus instead of computing numbers in those programs, the information needed, such as mean, standard deviation, and p-value, would be part of the ERT output.

From our assigned perspective, we would change the front end of GIFT to increase both usability and aesthetics. One example would be making use of the whole screen while giving surveys. There may have actually been a threat to internal validity by presenting a frustrating screen that lacked the space and screen usage that was allotted to the survey portion. The frustration derived from having to scroll to the right and left to read a question when the entire screen was virtually unused could have easily created skewed scores, especially the mood ratings. Another would be to take the logic puzzle out of Microsoft PowerPoint and put it on the HTML (hyper text markup language) display with the surveys, using radio buttons to mark choices instead of an "x" or "o".

Another suggestion would be to change the data output to column format, which would be easier to interpret. The most important change would be making GIFT more accessible. In summary, the user should be able to double-click an icon and start using the program. They should not have to open multiple files or wait for green lights. To make GIFT easier to use without a human tutor there needs to be some audio interaction. It does not have to be extensive, but small, simple audio cues would be more effective than the repetitive text boxes currently used. They would break up the sensory load and encourage users to be more attentive by introducing a different stimulus.

The way ahead for a program like GIFT lies in increased and simplified user interaction. This way ahead could be in the form an avatar that takes the place of the human in the tutoring process. Much like Apple's "Siri" or the Microsoft Word "Paperclip", this avatar could be used as a mascot for the program vehicle to present important and helpful messages to users in a way they are more familiar with. Along with being a guide through tutorials, it could also double as a device to answer frequently asked questions. The avatar entity could increase both the learning and retention effect of the program. Users and participants would have an icon to help them remember and identify the program. Through proper implementation, users may build a relationship with this entity within the program and become increasingly comfortable relying on

the avatar as they progress through various scenarios. It is important to remember that the absence of a human tutor requires simplicity for the average user to get through the program efficiently.

A final take-away is that human-less tutoring does not necessarily mean the program must be void of human-like interaction. In the short-term, this could look as simple as a small object that has been personified to guide the users. However, in the future this could result in the integration of GIFT and full-scale human modeling computer programs. The visual and auditory senses could be engaged while the user also received haptic feedback for physical tasks, thus making GIFT the ultimate human-less tutor. GIFT is the beginning of a long line of human-less tutors that could potentially change the way we train, the way we learn, and even the way we live.

4. References

Bloom BS. The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher. 1984;13(6):4–16.

Pavlik PI Jr, Brawner K, Olney A, Mitrovik A. A review of student models used in intelligent tutoring systems. In: Sottilare R, Graesser A, Hu X, Holden H, editors. Design recommendations for intelligent tutoring systems – volume 1: learner modeling. Orlando (FL): Army Research Laboratory (US); c2013. p. 39–40.

Sottilare, RA. Pushing and pulling toward future ITS learner modeling concepts. In: Sottilare R, Graesser A, Hu X, Holden H, editors. Design recommendations for intelligent tutoring systems; volume 1: learner modeling. Orlando (FL): Army Research Laboratory (US); c2013. p. 195–197.

Sottilare RA, Brawner KW, Goldberg BS, Holden HK. The generalized intelligent framework for tutoring (GIFT). Orlando (FL): Army Research Laboratory (US); 2012; concept paper.

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

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL IMAL HRA MAIL & RECORDS MGMT
1 (PDF)	GOVT PRINTG OFC A MALHOTRA
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRM D T DAVIS BLDG 5400 RM C242 REDSTONE ARSENAL AL 35898-7290
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRS EA DR V J RICE BLDG 4011 RM 217 1750 GREELEY RD FORT SAM HOUSTON TX 78234-5002
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRM DG J RUBINSTEIN BLDG 333 PICATINNY ARSENAL NJ 07806-5000
1 (PDF)	ARMY RSCH LABORATORY – HRED ARMC FIELD ELEMENT RDRL HRM CH C BURNS THIRD AVE BLDG 1467B RM 336 FORT KNOX KY 40121
1 (PDF)	ARMY RSCH LABORATORY – HRED AWC FIELD ELEMENT RDRL HRM DJ D DURBIN BLDG 4506 (DCD) RM 107 FORT RUCKER AL 36362-5000
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRM CK J REINHART 10125 KINGMAN RD BLDG 317 FORT BELVOIR VA 22060-5828
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRM AY M BARNES 2520 HEALY AVE STE 1172 BLDG 51005 FORT HUACHUCA AZ 85613-7069
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRM AP D UNGVARSKY POPE HALL BLDG 470 BCBL 806 HARRISON DR FORT LEAVENWORTH KS 66027-2302
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRM AT J CHEN 12423 RESEARCH PKWY ORLANDO FL 32826-3276
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRM AT C KORTENHAUS 12350 RESEARCH PKWY ORLANDO FL 32826-3276
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRM CU B LUTAS-SPENCER 6501 E 11 MILE RD MS 284 BLDG 200A 2ND FL RM 2104 WARREN MI 48397-5000
1 (PDF)	ARMY RSCH LABORATORY – HRED FIRES CTR OF EXCELLENCE FIELD ELEMENT RDRL HRM AF C HERNANDEZ 3040 NW AUSTIN RD RM 221 FORT SILL OK 73503-9043
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRM AV W CULBERTSON 91012 STATION AVE FORT HOOD TX 76544-5073
1 (PDF)	ARMY RSCH LABORATORY – HRED RDRL HRM DE A MARES 1733 PLEASONTON RD BOX 3 FORT BLISS TX 79916-6816
8 (PDF)	ARMY RSCH LABORATORY – HRED SIMULATION & TRAINING TECHNOLOGY CENTER RDRL HRT COL G LAASE RDRL HRT I MARTINEZ RDRL HRT T R SOTTILARE RDRL HRT B N FINKELSTEIN RDRL HRT G A RODRIGUEZ RDRL HRT I J HART RDRL HRT M C METEVIER RDRL HRT S B PETTIT 12423 RESEARCH PARKWAY ORLANDO FL 32826