Proceedings of the Seventh Annual GIFT Users Symposium

May 2019
Orlando, Florida

Edited by:
Benjamin S. Goldberg

Part of the Adaptive Tutoring Series

Proceedings of the 7th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym7)

Edited by:
Benjamin Goldberg

Printed in the United States of America
First Printing, May 2019

U.S. Army Combat Capabilities Development Command
Soldier Center
Orlando, Florida

International Standard Book Number: 978-0-9977257-6-6

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

CONTENTS

FROM THE EDITOR

GIFT is a free, modular, open-source tutoring architecture that is being developed to capture best tutoring practices and support rapid authoring, reuse and interoperability of Intelligent Tutoring Systems (ITSs). The authoring tools have been designed to lower costs and entry skills needed to author ITSs and our research continues to seek and discover ways to enhance the adaptiveness of ITSs to support self-regulated learning (SRL).

This year marks the seventh year of GIFT Symposia and we accepted 20 papers for publication. None of this could happen without the efforts of a fantastic team. Our program committee this year did an outstanding job organizing and reviewing, and we want to recognize them for their efforts.

- Michael Boyce
- Elyse Burmester
- Keith Brawner
- Jeanine DeFalco
- Ben Goldberg
- Greg Goodwin
- Michael Hoffman
- Joan Johnston
- Jong Kim
- Rodney Long
- Anne Sinatra

We are proud of what we have been able to accomplish with the help of our user community. This is the fifth year we have been able to capture the research and development efforts related to the Generalized Intelligent Framework for Tutoring (GIFT) community which at the writing of these proceedings has well over 1000 users in over 65 countries.

These proceedings are intended to document the evolutions of GIFT as a tool for the authoring of intelligent tutoring systems (ITSs) and the evaluation of adaptive instructional tools and methods. Papers in this volume were selected with the following goals in mind:

- The candidate papers describe tools and methods that raise the level of knowledge and/or capability in the ITS research and development community
- The candidate papers describe research, features, or practical applications of GIFT
- The candidate papers expand ITSs into previously untapped domains
- The candidate papers build/expand models of automated instruction for individuals and/or teams

The editors wish to thank each of the authors for their efforts in the development of the ideas detailed in their papers. As a community we continue to move forward in solving some significant challenges in the ITS world.

GIFT and the GIFT Symposium will take on a broader perspective as the new Center for Adaptive Instructional Sciences (CAIS) begins formal operations under ARL's Open Campus Initiative. The purpose of CAIS is to encourage the community development of adaptive instructional capabilities & standards. You can learn more about CAIS at https://www.arl.army.mil/opencampus/centers/cais.

Also new this year is GIFT Summer Camp which will pilot in June 2017. GIFT Summer Camp will teach an initial group of GIFT stakeholders how to author adaptive tutors using GIFT. Summer Camp follows on the heels of a successful assessment of the GIFT authoring tools earlier this year. Our intent is to open Summer Camp up to public users in 2018.

Finally, GIFT instructional videos will be available on YouTube this summer.

We would also like to encourage readers to follow GIFT news and publications at www.GIFTtutoring.org. In addition to our annual GIFTSym proceedings, GIFTtutoring.org also includes volumes of the Design Recommendations of Intelligent Tutoring Systems, technical reports, journal articles, and conference papers. GIFTtutoring.org also includes a users' forum to allow our community to provide feedback on GIFT and influence its future development.

Many thanks to all GIFT users…

Ben

Benjamin Goldberg, Ph.D.
GIFTSym7 Chair and Proceedings Editor

THEME I: GIFT OVERVIEW AND UTILITY

Architecture and Ontology in the Generalized Intelligent Framework for Tutoring: 2019 Update

Keith Brawner[1], Michael Hoffman[2], Benjamin Nye[3], Chris Meyer[4]
U.S. Army Research Laboratory[1], Dignitas Technologies[2], Institute for Creative Technologies University of Southern California[3], Synaptic Sparks Incorporated[4]

INTRODUCTION

The first version of the Generalized Intelligent Framework for Tutoring (GIFT) was released to the public in May of 2012. One year later, the first symposium of the GIFT user community was held at the Artificial Intelligence and Education conference in Memphis, Tennessee. Since then, the GIFT development team has continued to gather feedback from the community regarding recommendations on how the GIFT project can continue to meet the needs of the user community and beyond. This current paper continues the conversation with the GIFT user community in regards to the architectural "behind the scenes" work and how the GIFT project is addressing the user requirements suggested in the previous GIFTSym6 proceedings. The development team takes comments within the symposium seriously, and this paper serves to address requirements from prior years.

As a follow up to the "GIFT 2015 Report Card and State of the Project" (Brawner & Ososky, 2015), the GIFT 2016 Community Report (Ososky & Brawner, 2016), the GIFT 2017 Architecture Report (Brawner, Heylmun, & Hoffman, 2017), and the 2018 paper (Brawner & Hoffman, 2018) the feature requests and responses have been broken out among a number of papers, and into logical sections of this work. This paper discusses the ongoing architectural workings and changes in support of the various sets of projects. The number of projects which the GIFT overall projects is now around 30, which continues to represent a) the inability for significant direct support of any individual project and b) the relatively little support that individual projects need to be successful. GIFT generally works well enough to support research studies without direct developer guidance or specifically developed features.

The remainder of this paper discusses the requirements requested from the last GIFTSym, the developed functionality new to this year and the continuation of community dialogue in paper form.

WELCOME

First, to the new members of the GIFT community and new GIFT users – Welcome! There are a number of recommended resources that will help to orient you to this project and ecosystem. GIFT has come a long way since its original goals were defined in its description paper (Sottilare, Brawner, Goldberg, & Holden, 2012). First, we would encourage you to simply get started, as the tools and example courses have been designed to try to be as easy as possible for the creation of intelligent tutoring systems.

If you struggle with any individual aspect of the system, however, the team has produced short "how to" videos to try to help around the sticking points. There are now many such videos available on the GIFT YouTube channel, which is the first result if you search "Generalized Intelligent Framework for Tutoring Youtube" on Google. The YouTube videos have not been updated for the new release, however, the vast majority of the GIFT challenges and authoring has remained unchanged.

In addition to a Quick Start Guide, usable tools, and videos, there is support for developers in the help forums and documentation. The GIFT user community is also invited to ask questions and share your experiences and feedback on our forums (https://gifttutoring.org/projects/gift/boards). The forums are actively monitored by a

small team of developers, in addition to a series of Government project managers. The forums are a reliable way to interact with the development team and other members of the GIFT com- munity. The forums, at the time of this writing, have over 1200 postings and responses. Documentation has been made freely available online at https://gifttutoring.org/projects/gift/wiki/Documentation, with interface control documentation https://gifttutoring.org/projects/gift/wiki/Interface_Control_Document_2018-1, and a developer guide https://gifttutoring.org/projects/gift/wiki/Developer_Guide_2018-1. These documents are updated each software release.

GIFT CLOUD GENERAL REPORTING

GIFT Cloud is now legacy, and has been running more-or-less continuously for the last three years, in support of numerous experiments. At the time of writing, it appears that only developers download the downloadable versions of GIFT. Cloud GIFT is kept online and updated in advance of the downloadable version, meaning that content must be backwards-ported to be compatible with the perpetually out of date offline version. We do our best to keep the downloadable version to regularly scheduled improvements, but, for ordinary users, we would encourage you to use the Cloud version – it is better supported and more stable than the downloadable version. It supports hundreds of simultaneous users forexperiments. Further, there are approximately 8 cloned cloud versions with different software configurations live at any given time. We are generally confident in the systems' ability to stay up and cope with demand.

Behind the scenes, however, the re-tooling to move to a deployment version of dev-desk to dev-cloud to production has been working well. The team has greater ability to bug requests, with faster turnaround time. In this paper we reiterate that a clone of cloud.gifttutoring.org is always available upon request, and we have granted several requests over the year – necessitating an updating of the instructions to deploy a new cloud build and the hardening of those instructions. The previous version of this paper (Brawner & Hoffman, 2018) identified a number of organizations which had requested special access, but this number is now too great to count individually.

Virtual Machines Available Upon Request

As part of the move to Cloud GIFT, we have a number of specialized processes which run in the back- ground. Figure 1 shows the current structure of the Virtual Machine (VM) instances which operate Cloud GIFT. At its basic level, GIFT runs on two VMs; a Windows VM for all of the core GIFT features, and a Linux VM hooked up to an Amazon Relational Database Service (RDS) for the content. These items are what are contained in the downloadable GIFT instance. In addition to the basic instances, however, are services for monitoring GIFT; PiWik monitors user behaviors within the system, while the GIFT monitoring service monitors usage for future performance improvements. GIFT now includes an instance to a Social Media Framework (SMF) and Learner Record Store (LRS), which are based around Elgg and Learning Locker, respectively. GIFT's copies of these configurable items are available upon request, and posted to github, but the authors would urge users to select their own instances of commercial sharing and data warehousing items dependent upon their own individual needs; there is nothing tying GIFT to a specific SMF, LRS, PiWik, or monitoring framework. We do not think of these items as core to GIFT, only that they are reported outwards.

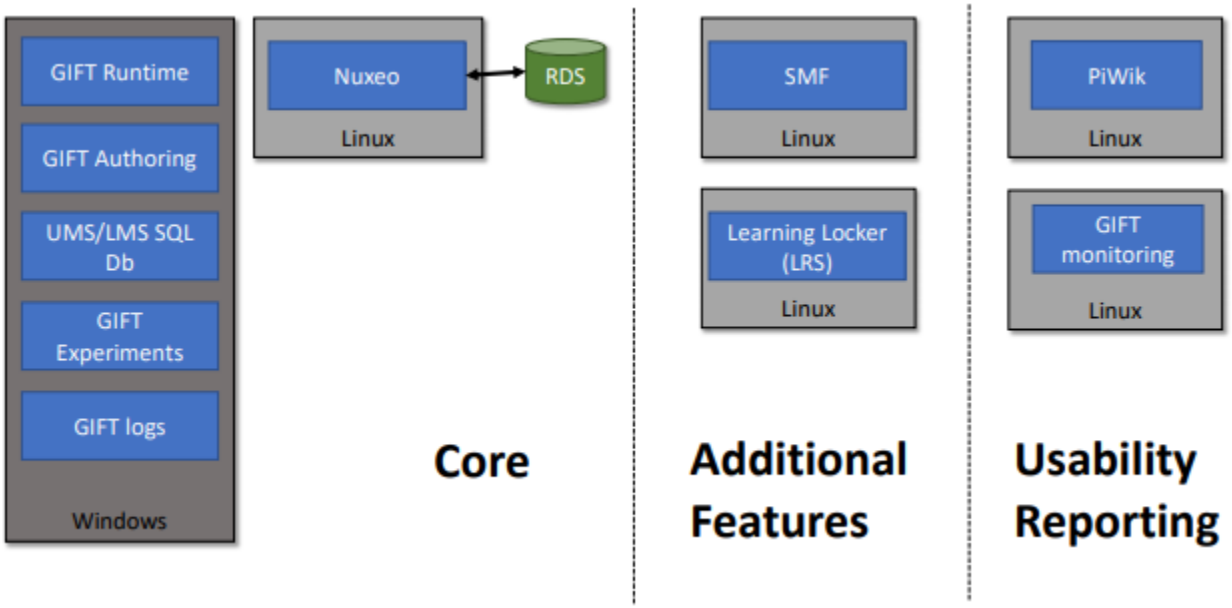


Figure 1: Simplistic Diagram of Cloud Gift Items

NEW INSTRUCTIONAL MODELS

A new Adaptive courseflow course object version was introduced in the interim release of GIFT, GIFT 2017-12-22. In this full release all legacy Adaptive courseflow course objects are now automatically converted to this new instance. Authors will now see a new icon for these course objects as well as an expanded Adaptive courseflow course object editor in the course creator. Learners taking a course with a legacy adaptive courseflow will mainly see a difference when it comes to remediation. In the past if you were deemed a novice on a course concept after a check on learning phase of an adaptive courseflow course object you would see both rule and example content as part of remediation. Now you will see a single piece of content for remediation of that concept. This is a part of upgrading to the reinforcement learning frame- work based upon the idea of Interactive-Constructive-Active-Passive methods of instruction (Rowe et al., 2018).

VIRTUAL HUMAN TOOLKIT (VHTK)

GIFT now supports 2 character servers, Media Semantics and Virtual Human. Both are available on the Downloads tab of gifttutoring.org. GIFT is now configured to use Virtual Human as the default character server, however you can still use custom Media Semantics characters in courses with no GIFT configuration changes by including the custom character in your course folder and then referencing that custom character appropriately (see Excavator and Explicit Feedback courses as examples). Note that if you need to run a character in IE 11 (or earlier) than you will need to use Media Semantics because Virtual Human uses the Unity WebGL player, which is not supported in older IE browsers. The development in this category was previously reported in a prior paper (Nye, Auerbach, Mehta, & Hartholt, 2017), but it is currently tested, released, and live. An additional two characters have been added to a development branch and are expected in future releases, assuming successful testing and validation.

LEARNING TOOLS INTEROPERABILITY (LTI)

Previous developments to the LTI interface was reported last year (CITE). This year involved minor tweaks of the interface, bilateral course sharing, and two course sharing publishing items. The LTI v2.0 interface was not particularly embraced by the community of developers at EdX, and thus the GIFT developers are following the lead of the larger Massive Online Open Courseware providers (Aleven et al., 2017) in moving to the updated v1.1.1.

LEARNER RECORD STORES AND COMPETENCIES

The authors wish to make the community aware that we are in the midst of integrating the Competency And Skills System (CASS) developed by the Advanced Distributed Learning (ADL) Initiative. The military and Army community have need of the technology represented by this community. Further, the xAPI community is embracing the technology through the implementation of xAPI Profiles (Bowe & Silvers, 2018) within the IEEE Learning Technology Standards Committee (LTSC) (Robson & Barr, 2018) In brief, the xAPI from GIFT courses informs competency assessments informs readiness assessments informs course recommendation which generates xAPI data in a virtuous cycle. We welcome participation, and more information on the exact developments can be found at:

- CASS - https://www.cassproject.org/
- xAPI Profiles - http://sites.ieee.org/sagroups-9274-1-1/
- LTSC - http://sites.ieee.org/sagroups-ltsc/home/

AUTHORING

Massive improvements in authoring have been made since the last release, through the integration of pre- vious versions of “GIFT Wrap” as well as the larger number of deployed courses recently. The process for authoring assessments of this kind needed to be streamlined, and has. This is especially true in regards to the authoring of a “Domain Knowledge File” (DKF) or “Real Time Assessment”, which is what it is called on Cloud GIFT. The improvements here can be captured in the below screenshot, showing the old and new interfaces side to side. The improvements are substantial and noticeable.

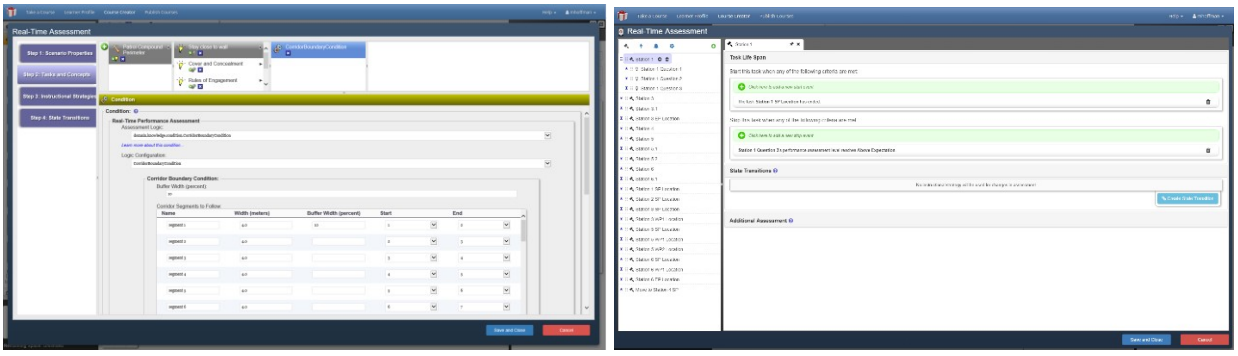


Figure 2: Authoring Tools changes for GIFT Wrap

RESEARCH DIRECTIONS: TEAM AND PSYCHOMOTOR TRAINING

Part of the goal of the GIFT project is to expand tutoring systems from relatively well-defined domains to ill-defined domains, from desktop training to "in the wild" training, and from individual training to team training. This is part of the military interest in intelligent tutoring technologies – Warfighters train as a group, and within the training environment. This section provides an update to last years' status on team training and psychomotor training.

Team Training

While specific research implementations can be read elsewhere within prior proceedings (Sinatra, 2018), the team has done specific work in order to show relevance to team training items. This technology was further developed into a demonstration at ITSEC and is now available for multiple projects upon request. This technology is scheduled for early implementation in the coming release, considering the priority that the Synthetic Training Environment (STE) is placing on team training, a "train as you fight" model, and on "25 bloodless battles" (Defense News, 2018).

Psychomotor Training

Psychomotor, or "in the wild" training is a significant part of the reason for military investments in the intelligent tutoring technologies. The prototype land navigation mobile application reported upon last year (CITE) has now been released as software functionality, with the gains made in the authoring tools placed into the current release. The developments in land navigation have significantly shaped the outcomes of the authoring tools for new functionality.

OTHER COMMUNITY-REQUESTED FEATURES

Wheelspinning Prediction

A request for a prediction of wheelspinning behavior was included in last years' proceedings (Park & Matsuda, 2018). The authors would hope that this implementation would be as simple as a sensor module plugin, taking the student answers as the input, doing processing in an interface of the author's choice (RapidMiner, Python, XML-RPC, etc.), and exporting a new student state on the concept level; e.g. "Con- cept 1":"Wheelspinning". At that point, the student state could be given an instructional remediation using the standard body of GIFT authoring tools from a Domain Knowledge File (RealTime Assessment). This type of implementation should be relatively straightforward, and is covered in the "how to add a sensor" portion of the online documentation, at: https://gifttutoring.org/projects/gift/wiki/Developer_Guide_2019- 1#Integrate-a-Sensor.

Validated Motivational Assessments

Motivational assessment, as requested by the UCF group in last year's GIFTSym is anticipated to transition into public use inside of the next 3 months (Biddle, Lameier, Reinerman-Jones, Matthews, & Boyce, 2018).

Natural Language Processing for Team Interactions

Efforts to begin researching natural language processing for the determination of team dynamics have be- gun, but have not yet made it into production. The authors welcome solutions and ideas which target language as a

manner of team assessment, as they have been specifically requested by the community of GIFT users (Johnston, 2018).

Human In-The-Loop Functionality

Recent projects, especially in the realm of larger teams, have demonstrated the need for a human-in-the- loop capability (McCormack, Kilcullen, Sinatra, Brown, & Beaubien, 2018). The idea behind this capability is that there will be an "auto mode" and a "manual mode" which function similar to a traditional intelligent tutoring system, and a ITS-as-you-approve-it feature set. Further, the human will be able to introduce manual assessments of any type that GIFT was initially programmed for. This functionality, as shown below, is currently available upon request, and is in testing internally on larger-scaled operations.

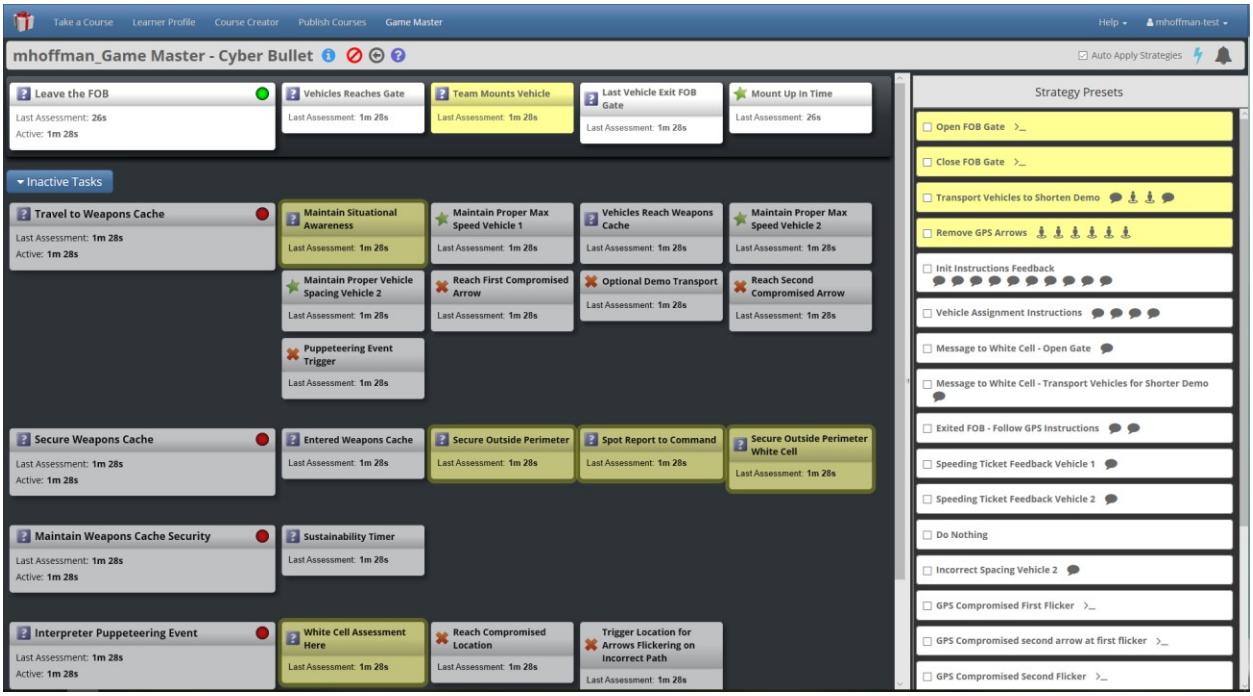


Figure 3: Enhanced Human-in-the-loop interface

Better User Guide

A better user guide was requested at last years' GIFTSym (Julian, 2018). The authors hope that the updated manuals, new material, and YouTube video series would be helpful.

Predictive Analysis of Performance and Training ROI computations

This request has gone unaddressed and we welcome others in the community to take it up.

GIFT AND IEEE STANDARDS

As part of last year's GIFT Symposium, there is an associated standards meeting. This standards meeting will be among those which occurred over the course of the year, including telephone calls, in-person meetings, proceedings presentations, and other activities. The IEEE Learning Technologies StandardsCommit- tee, with support from the GIFT community and the Government, is now seeking involvement in standardization

activities. The GIFT community invites the reader to join the conversation on what data exchange standards for learning technologies might look like in the future – there is now active IEEE community on the subject, to which the GIFT project is contributing meaningfully. Interested readers are encouraged to go to the IEEE LTSC meetings to become involved.

REFERENCES

Aleven, V., Baker, R., Blomberg, N., Andres, J. M., Sewall, J., Wang, Y., & Popescu, O. (2017). *Integrating MOOCs and Intelligent Tutoring Systems: edX, GIFT, and CTAT.* Paper presented at the Proceedings of the 5th Annual Generalized Intelligent Framework for Tutoring Users Symposium, Orlando, FL, USA.

Biddle, E., Lameier, E., Reinerman-Jones, L., Matthews, G., & Boyce, M. (2018). *Personality: A key to Motivating our Learners.* Paper presented at the 6th Annual GIFT Users Symposium.

Bowe, M., & Silvers, A. E. (2018). US DoD xAPI Profile Server Recommendations.

Brawner, K., Heylmun, Z., & Hoffman, M. (2017). *The GIFT 2017 Architecture Report.* Paper presented at the Proceedings of the 5th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym5).

Brawner, K., & Hoffman, M. (2018). *Architecture and Ontology in the Generalized Intelligent Framework for Tutoring: 2018 Update.* Paper presented at the Proceedings of the 6th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym6).

Brawner, K., & Ososky, S. (2015). *The GIFT 2015 Report Card and the State of the Project.* Paper presented at the Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym3), Orlando, FL.

Defense News, M. G. (Producer). (2018). 25 bloodless battles: Synthetic training will help prepare for current and future operations. Retrieved from https://www.defensenews.com/smr/defense-news-conference/2018/09/05/25-bloodless-battles-synthetic-training-will-help-prepare-for-current-and-future-operations/

Johnston, J. H. (2018). *Team Performance and Assessment in GIFT–Research recommendations based on Lessons Learned from the Squad Overmatch Research Program.* Paper presented at the Proceedings of the Sixth Annual GIFT Users Symposium.

Julian, D. (2018). *Final Report for IDS6938 – Intelligent Tutoring System Design: Basic Robotic Course.* Paper presented at the Proceedings of the Sixth Annual GIFT Users Symposium.

McCormack, R. K., Kilcullen, T., Sinatra, A. M., Brown, T., & Beaubien, J. M. (2018). *Scenarios for training teamwork skills in virtual environments with GIFT.* Paper presented at the Proceedings of the Sixth Annual GIFT Users Symposium.

Ososky, S., & Brawner, K. (2016). *The GIFT 2016 community report.* Paper presented at the Proceedings of the 4th Annual GIFT Users Symposium.

Park, S., & Matsuda, N. (2018). *Predicting Students' Unproductive Failure on Intelligent Tutors in Adaptive Online Courseware.* Paper presented at the Proceedings of the Sixth Annual GIFT Users Symposium.

Robson, R., & Barr, A. (2018). *Learning Technology Standards-the New Awakening.* Paper presented at the Proceedings of the Sixth Annual GIFT Users Symposium.

Rowe, J., Spain, R., Pokorny, B., Mott, B., Goldberg, B., & Lester, J. (2018). *Design and Development of an Adaptive Hypermedia-Based Course for Counterinsurgency Training in GIFT: Opportunities and Lessons Learned.* Paper presented at the proceedings of 6th Annual GIFT Users Symposium.

Sinatra, A. (2018). *Team Models in the Generalized Intelligent Framework for Tutoring: 2018 Update.* Paper presented at the Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym6), Orlando, FL.

Sottilare, R., Brawner, K. W., Goldberg, B. S., & Holden, H. A. (2012). The Generalized Intelligent Framework for Tutoring (GIFT).

ABOUT THE AUTHORS

Keith Brawner, PhD *is a senior researcher for the U. S. Army Combat Capability Development Command Soldier Center at the Simulation and Training Technology Center (CCDC-SC-STTC), and is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT). He has 13 years of experience within U.S. Army and Navy acquisition, development, and research agencies. He holds a Masters and PhD degree in Computer Engineering with a focus onIntelligent Systems and Machine Learning from the University of Central Florida. His current efforts are on artificial intelligence for the Synthetic Training Environment Simulation and Network Compression. He manages research in adaptive training, semi/fully automated user tools for adaptive training content, and architectural programs towards next-generation training.*

Michael Hoffman *is a senior software engineer at Dignitas Technologies and the technical lead for the GIFT project. He has been responsible for ensuring that the development of GIFT, meeting community requirements, and supporting production ITS systems, ITS research, and the growing user community. Michael manages and contributes support for the GIFT community through various mediums including the GIFT portal (www.GIFTTutoring.org), annual GIFT Symposium conferences and technical exchanges with ARL and their contractors. In addition he utilizes his expertise in integrating third party capabilities such as software and hardware systems to enable other organizations to inte- grate GIFT into their training solutions.*

Benjamin Nye, Ph.D*. is the Director of Learning Science at the University of Southern California, Institute of Creative Technologies (USC-ICT). He will lead the project's management and execution, including coordinating software efforts, mentor interviews, study design, and data collection. This leadership will involve leading efforts from specialists from in-house ICT design teams, art teams, and learning science teams to accomplish these roles. Dr. Nye's research has been recognized for excellence in intelligent tutoring systems (1rst Place ONR ITS STEM Grand Challenge; Nye, Windsor, et al., 2015; Nye et al., 2014), cognitive agents (BRIMS 2012 best paper; Nye & Silverman, 2013; Nye, 2012), and realistic behavior in training simulations (Federal Virtual Worlds Challenge; Silverman et al., 2012). His research is on scalable learning technologies (Nye et al., 2014) and design principles that promote learning (Nye, Graesser, & Hu, 2014; Nye, 2014; Nye, Morrison, & Samei, 2015). This research has led to 20 peer-reviewed papers, 11 book chapters, and 5 open-source projects.*

Christopher Meyer *brings a breadth of leadership experience and technical knowledge to the team. And, most recently, Christopher has supported the GIFT program for two years under the most current contract. He received his Bachelor and Master of Science degrees in Computer Science from Kansas State University, also receiving minors in Economics and Modern Languages, and studied abroad for a year during a tour in Japan at Chukyo University dedicated to the specialized study of Artificial Intelligence. After completing traditional education phases, Chris was employed at Lockheed Martin for 10 years working hand-in-hand with representatives from the Departments of Defense, Health and Human Services, Energy, and Education to assist in the creation of solutions to solve challenges at a national level. Having now co-created his own business segment, Chris enjoys utilizing entrepreneurship, international experience, leadership knowledge, and his own engineering skills alongside his peers to advance world technology, health, and opportunity efficiently and responsibly.*

The 2019 Instructor's Guide to GIFT

Anne M. Sinatra
U.S. Army Combat Capabilities Development Command (CCDC) – Soldier Center – Simulation and Training Technology Center (STTC)

Introduction

The Generalized Intelligent Framework for Tutoring (GIFT) is a research project which is continually developing (Sottilare, Brawner, Sinatra, & Johnston, 2017). As a result of this, many of the functionalities within GIFT continue to expand, and some of the interfaces that users interact with are being updated as well. I began writing a series of "Research Psychologist's Guides" for GIFT in 2014 (Sinatra, 2014; Sinatra, 2016; Sinatra, 2018), and have updated them on a bi-annual basis. The updates to the guide have captured many of the continuing changes that have occurred in GIFT, with the authoring tool and survey authoring systems being dramatically different between the publication times. As a companion piece to these guides, in 2015 the first Instructors' Guide to GIFT was published (Sinatra, 2015). This guide was written specifically from the perspective of an Instructor who might incorporate GIFT into the class that they were teaching in a number of ways: as an interactive medium for students to learn materials through, as a way to perform assessments, and even as a way to have students create projects for their classes. Since the publication in 2015 there have been many updates to GIFT including GIFT Cloud. The introduction of GIFT Cloud and access through the internet greatly increases the opportunities that an instructor has in using GIFT in their courses. One of the largest barriers to implementing GIFT in the classroom at the time of writing of the original guide was that both the instructor and students would need to install the GIFT soft- ware. Due to the implementation of GIFT Cloud, there is no longer a need to install the software, and the majority of GIFT's functions can be accessed from a web-enabled computer. This is a very large change and update to GIFT which influences the way that an instructor may use GIFT in a course. Following the format of the "Research Psychologist's Guide" to GIFT, the current work is an update to the original which also discusses the improvements that have been made to GIFT and new strategies to use with it since the last publication.

This current Instructors' Guide to GIFT expands upon the original instructor's guide, and also discusses ways that GIFT Cloud can be implemented in a class. In the current paper, there is an explanation of GIFT's tools from the perspective of an instructor, and an explanation of how to add previously existing content (e.g., exam questions; test banks; PowerPoint slides) to GIFT. While there are other documents which de- scribe using GIFT's tools, the current guide specifically discusses how to leverage the tools within GIFT to specifically from the perspective of an instructor who is concerned with grades and content in a formal class. An explanation is provided about how to extract data after a student has interacted with the course. There is additional discussion of the current state of GIFT and improvements that could be made in order to make improve its functionality for instructors.

USING GIFT IN A CLASS

GIFT can be used to create materials that students interact with either in person (in a computer lab) or on their own time. These materials can be used as a primary means of providing information (e.g., in an online course), or as an opportunity to review material on the student's own time. It is up to the instructor to decide how he or she would like to implement GIFT as part of a class. In the current guide there is discussion of the current functionality that exists, which an instructor can use to decide how to implement GIFT in his or her class. The most straightforward way for an instructor to use GIFT in a class is by creating a linear GIFT course and assigning it to students through a link. The current document discusses the tools that are relevant for creating a linear GIFT course and publishing it for distribution to students. The remediation and adaptive tutoring functionalities of GIFT may be of interest to advanced users, but are beyond the scope of the current guide. Additional information on how to use these functions can be found on the GIFT YouTube channel.

With GIFT, an instructor can create a number of different "GIFT Courses" which students can interact with. Depending on the author of the courses, each course can be on an individual topic (similar to a module), or they can include multiple learning objectives. If an instructor wishes to create a non-adaptive, linear course which utilizes pre-determined surveys then advanced features such as defining concepts do not need to be utilized within GIFT. However, if the course instructor wishes to implement remediation/adaptive course flow and/or utilize the Question Bank feature of GIFT for randomized questions, he or she will need to identify course concepts within the GIFT authoring tool. These course concepts can then be linked to the specific items for remediation and the individual questions that are authored in the course specific Question Bank.

Regardless of the specific way that an instructor plans to implement GIFT within his or her class, they could create materials that support each of the lessons that they are teaching in class and can implement them in the form of GIFT courses. These courses could then either be assigned as optional or required assignments that students can complete on their own time. Additionally, based on the preference of the instructor these materials could either be used for self-regulated review, or as actual graded assignments. In the case of using GIFT for graded assignments, there would be the additional question that the instructor would need to answer – would they be grading for completion of the assignment or actually grading based on the answers and activities that the student performs during the GIFT interaction. While both of these options are possible they would require different actions to be taken by the instructor to ensure that the relevant information is provided in order for them to get the information that they need.

USING GIFT FOR STUDENT ASSIGNMENTS

As mentioned above, GIFT can be used to provide materials and assignments to students in the form of interactions and quizzes that can be used for grades. Additionally, it can be used as a means for presenting materials to students either in class in a computer lab format, or on their own time. One application of GIFT that has been used previously, is to have students interact with GIFT and create their own intelligent tutoring systems (ITSs). Students can be assigned a specific topic that they need to create materials about, and then tasked to create their own ITS with GIFT. Versions of this assignment have been used with students of varying backgrounds and varying education levels including both undergraduate and graduate level. Additionally, if students wish to create their own research projects they can leverage GIFT as a means to do so. This type of assignment may be of particular interest in the field of human computer interaction, or ITS classes. Additionally, since GIFT is an on-going research project, students who complete usability assessments of GIFT could submit their outcomes and suggestions for consideration for possible future updates to the overall system.

GIFT FEATURES THAT ARE USEFUL FOR INSTRUCTORS

There are many tools and features of GIFT that are of interest to instructors who wish to implement GIFT within their classes. The most important items are the GIFT Authoring Tool, the new Survey Authoring System, and the "Publish Course" functionality.

GIFT Authoring Tool

The GIFT Authoring tool has gone through many updates and iterations through the years, and currently features an easy to use drag and drop interface. The left side of tool offers the possible course objects that can be utilized, and the right side of the tool is a course map that displays the order of the objects within the courseflow. Once an object is dragged from the left side of the screen to the main courseflow, it can be authored. The first action is to name the object. This name will be helpful for the instructor so that he or she knows what it is called when they are looking at their overall courseflow. After naming it on the right side of the screen a properties panel will appear that allows for customization of the object. Once an object is authored it can be

reordered in the courseflow by clicking on the object with the left mouse button and dragging it to a different place. See Figure 1 for screenshot example of the interface which shows a high- lighted object.

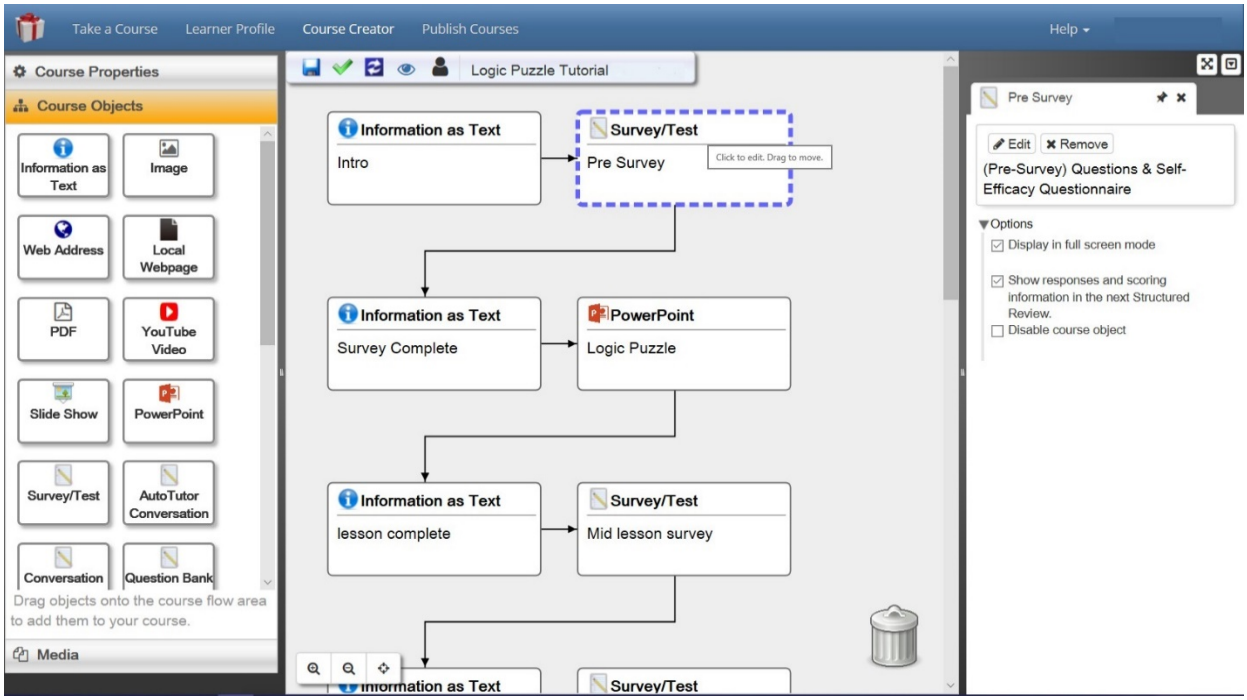


Figure 1. Screenshot of the GIFT Authoring Tool interface with a Survey/Test item highlighted. Course objects are on the left side of the screen, the courseflow is in the center, and the right side of the screen has specific course object properties.

The course objects that are available for use include Information as Text, Image, Web Address, Local Webpage, PDF, YouTube video, Slide Show, PowerPoint, Survey/Test, Conversation, Question Bank, Adaptive Courseflow (this is an advanced feature, and is beyond the scope of the current paper), Structured Review, and external connections with programs such as Virtual Battlespace. In many cases when amedia item is selected (e.g., image) the instructor can find it locally on his or her computer and it will be uploaded to their specific GIFT Cloud account.

Course Objects

If the instructor wants to provide information to the students he or she can use "Information as Text" or create an .html file that can be uploaded as a local webpage. If the instructor wants to send students to an external website, the Web Address can be used. It will bring up the webpage with a "Continue" button centered at the bottom of the GIFT interface. This can sometimes lead to potential student error if they click the "Continue" button before reading the webpage, so it can be helpful to include an "Information as Text" object prior to this which explains what they need to do in order to engage with the webpage.

Slide Show Object and PowerPoint Object

Two objects that are of particular note to instructors, and have very different implications for the way that students will interact with the course are "Slide Show" and "PowerPoint". Both of these course objects are created from a PowerPoint show file (.pps) that the instructor uploads to the course. However, if the PowerPoint that already exists consists of only text and static images then the preferred method to use is the "Slide Show" course object. When using the "Slide Show" course object, GIFT will convert the existing PowerPoint show file into images, which students will be prompted to read and advance through. There are a number of different options in order to adjust the interface as the instructor wishes to reduce the chance that a student will

accidentally skip through the slides. By using the "Slide Show" course object it allows for students to view the material without needing to download anything to their computers. This creates a fluid online experience.

The only instances where one would want to use a "PowerPoint" object file type is if there are macros or videos embedded in the existing original PowerPoint file which are vital to the content. In order to use the "PowerPoint" course object, GIFT will need to connect with an instance of PowerPoint that is on the student's computer. This means that in order to run the course the student will need to have a compatible version of PowerPoint installed on their computer. Additionally, it will require the student to download the gateway module that connects GIFT Cloud to their version of PowerPoint on the computer. This can lead to user error, or difficulty with running the specific course. Therefore, it is preferable to use the "Slide Show" object whenever possible. Existing PowerPoints can be saved as .pps files and they will be automatically converted to images for the instructor when using the "Slide Show" object. See Figure 2 for how to save your PowerPoint document as PowerPoint Show (.pps), which can then be used by GIFT.

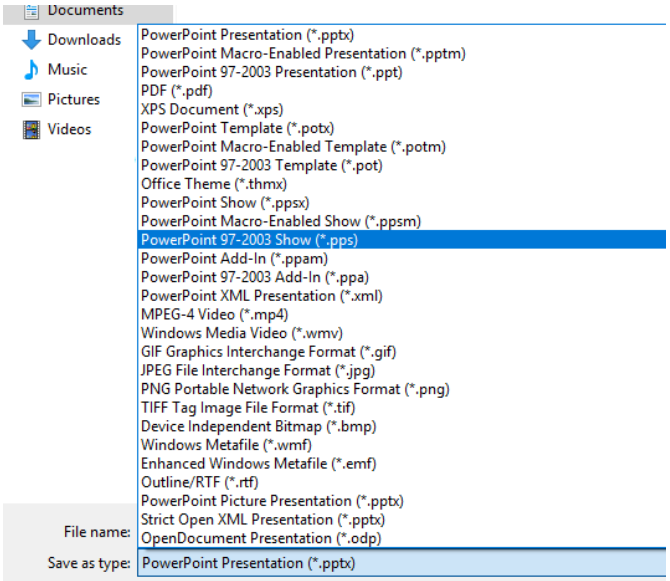


Figure 2. The correct format to save your PowerPoint as for use with GIFT as a Slide Show or PowerPoint object is "PowerPoint 97-2003 Show (*.pps) as highlighted above.

Adding Course Concepts

If an instructor wants to teach more than one concept in a GIFT course, or wants to use Adaptive Courseflow or the Question Bank object, then it is necessary to define Course Concepts in GIFT. To do so on the GIFT Authoring Tool, click on "Course Properties". Then click on "Concepts". Figure 3 shows the correct item to click on, and Figure 4 shows what the concept interface looks like. Multiple concepts can be created, and these will later be used to both tag and identify questions in the system that are associated with the proper concepts.

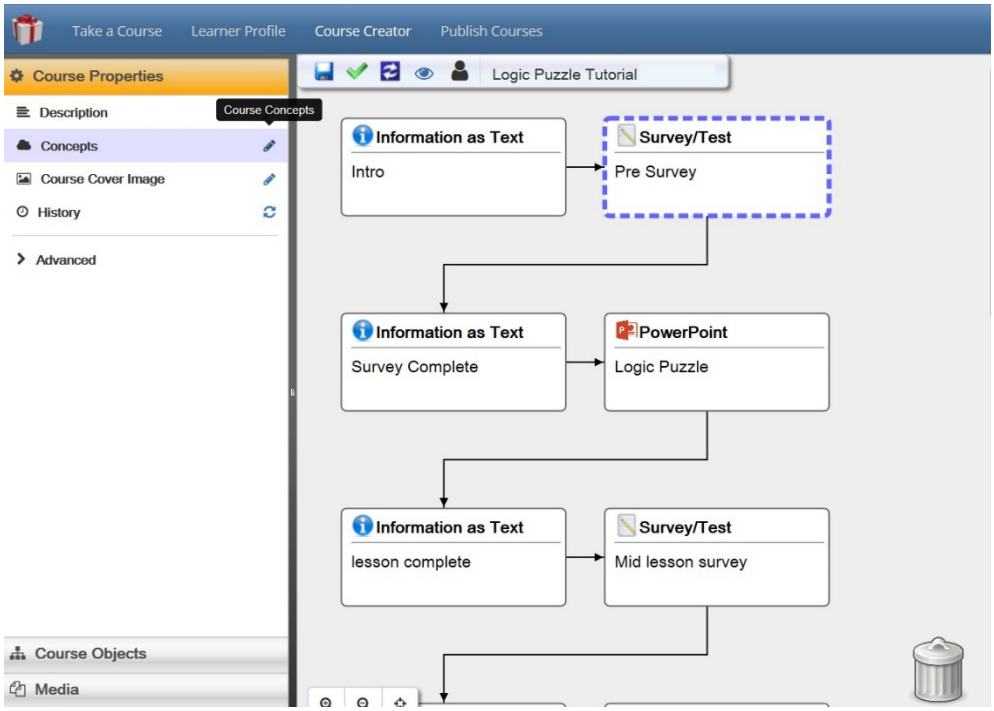


Figure 3. To define Concepts for a GIFT Course click on Course Properties, then the pencil next to Concepts.

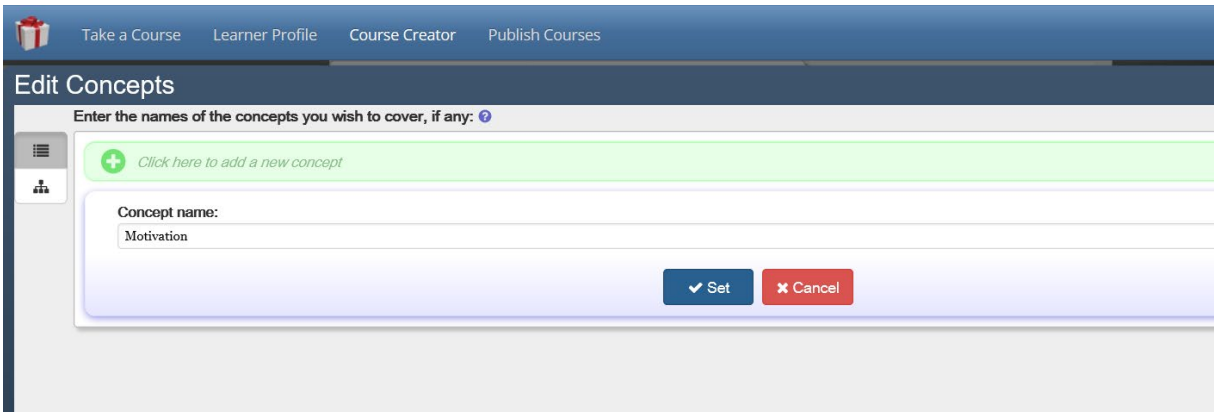


Figure 4. Once "Concepts" has been clicked on the interface in this figure will be displayed. For each of your concepts add it by clicking on the green plus button, and then give it a name. By adding these concepts they can later be used in the course.

Question Bank and Survey Course Objects

There is a distinction between using the Question Bank and Survey course objects in GIFT. If an instructor always wants to create a formal quiz or exam in which the same questions display in every instance, then the Survey course object should be used. If questions associated with specific concepts should be selected at random from a bank of questions of varying degrees of difficulty, with different concepts identified, then the Question Bank should be used.

Survey Course Object

Once a Survey object is selected, an additional selection needs to be made about the type of survey. A different option will be selected if the survey is actionable, non-actionable (the information will not be calculated in real-time for use), or an assessment of learner knowledge. In most cases within a class it would be expected that non-actionable information would be used when collecting straightforward information from the student such as their name, and the assess learner information selection would be used if the items have correct answers and will be automatically graded by the system. See Figure 5 for a screen- shot of the "Assess Learner Knowledge" option. Note that there is both a writing mode, and a scoring mode that can be selected from the top middle of the interface. It is also important to add a "Tag" to the Question Properties for each question. This will be the name that is available at the top of the column when the data is extracted.

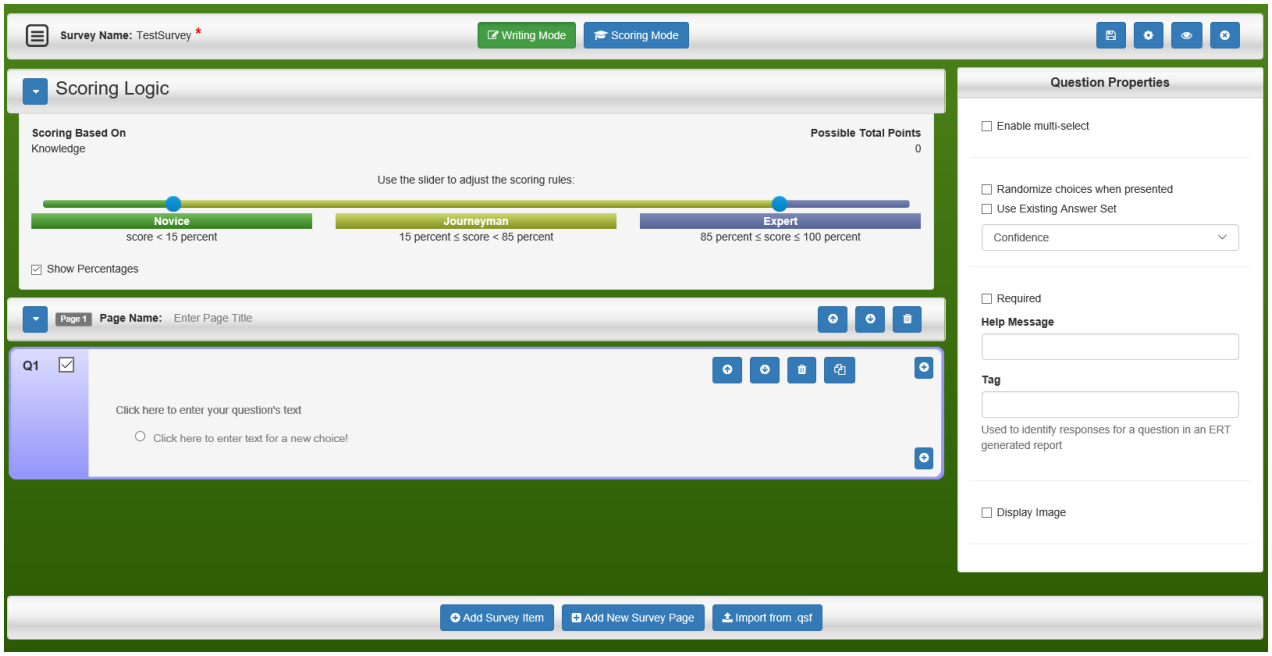


Figure 5. Screenshot of the Survey Interface

Question Bank Object

Each course has its own question bank which is directly linked to concepts that have been defined within the course. Once the Question Bank course object is added, there are two actions that need to be taken. First, questions need to be added to the overall course question bank. Second, the specific questions and concepts that will be assessed in the Question Bank object need to be selected. The number and types of questions will also be identified, as well as the number of correct questions to fall into each type knowledge level. An example of the initial configuration interface for the Question Bank object is in Figure 6. Note that the first step is to click on the "Course Question Bank" item on the top of the screen, which will then take the instructor to the question bank interface.

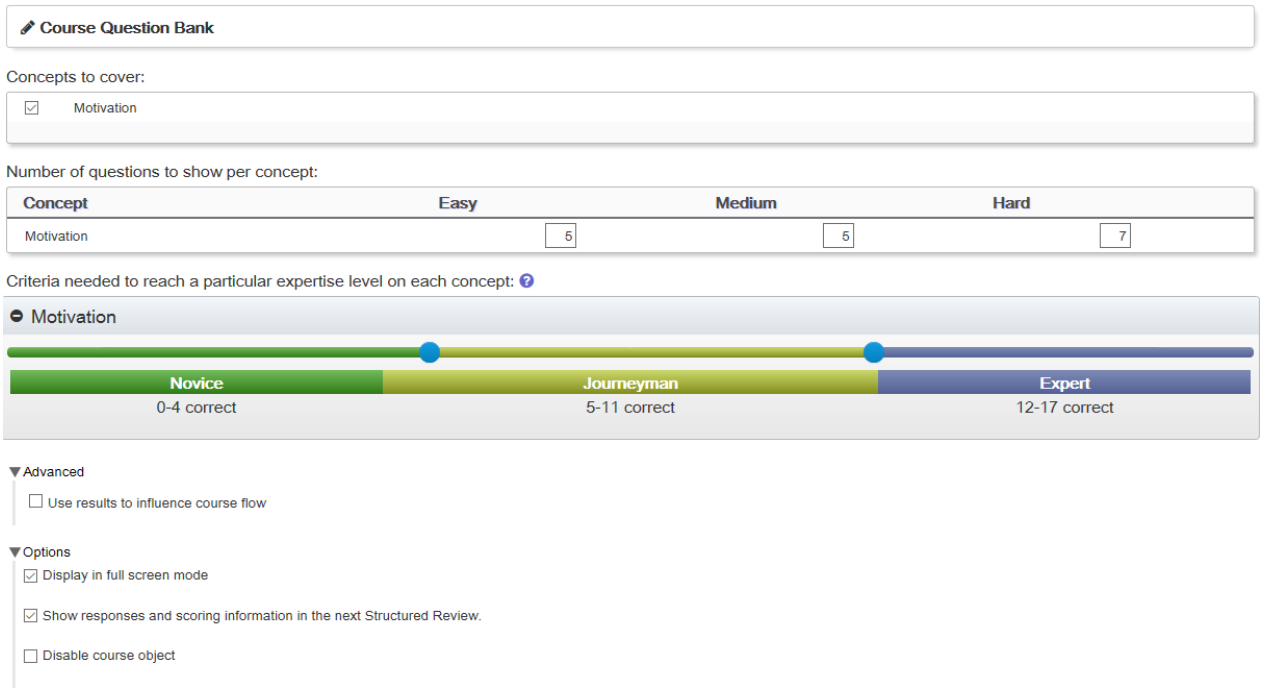


Figure 6. Question Bank object interface. First click on "Course Question Bank" to start adding questions to the overall bank for the course. Next, select the course concepts that will be identified with this specific in- stance of the Question Bank object within your courseflow.

Additionally, the questions that are entered in the main question bank are separate from those that are avail- able in the Survey object discussed previously. The interfaces are very similar, but in order to accurately associate the question bank item with a concept the "Scoring Mode" button needs to be clicked, and the "Question Difficulty" and "Associated Concepts" selected. See Figure 7 for an example.

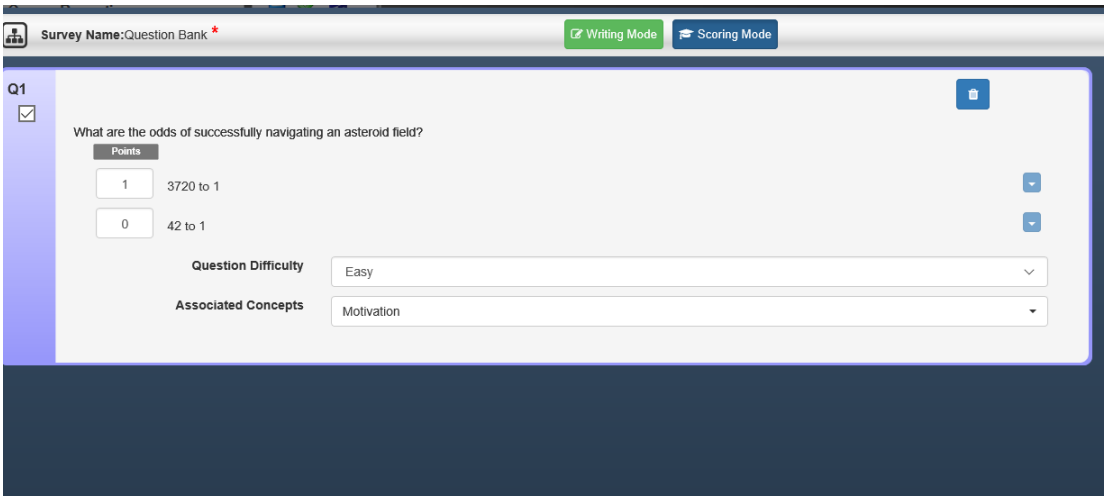


Figure 7. An example of Scoring Mode in a Question Bank. It is important to set the Question Difficulty, and select associated concepts.

Publish Course and Data Extraction

There are two methods that can be used to provide a course to an individual student. These include either exporting the course for import to the student's GIFT Cloud account, or publishing the course. The recommended method to use at the current time is the "Publish Course" functionality. At current time, while importing the course would result in linking the scores to an individual student's account, there is no way for the instructor to automatically retrieve the information that students enter in the system. This information is stored on the main instance of GIFT running on the Cloud and would require the GIFT team to provide it. The current solution is to use the "Publish Course" option. Publish Course takes the existing version of a course and provides a URL for it so that students can access it from the link. Since this method is not directly linked to a student's GIFT account, in order to use the output for grade purposes, a demographic question will need to be added to the course that asks that student to enter his or her name or a relevant student ID number.

Once "Publish Course" is selected it brings up an interface that has red or green bars for already existing published courses. To publish a new course, click the "Publish Course" button on the top left side of the screen. While the overall terminology has been updated, this functionality was originally used for experiments, and the correct selection to make on the pop up screen is "Publish Course as Experiment". You then type in a name for the course, and then select a course from the displayed list. See Figure 8 for a screenshot of this interface.

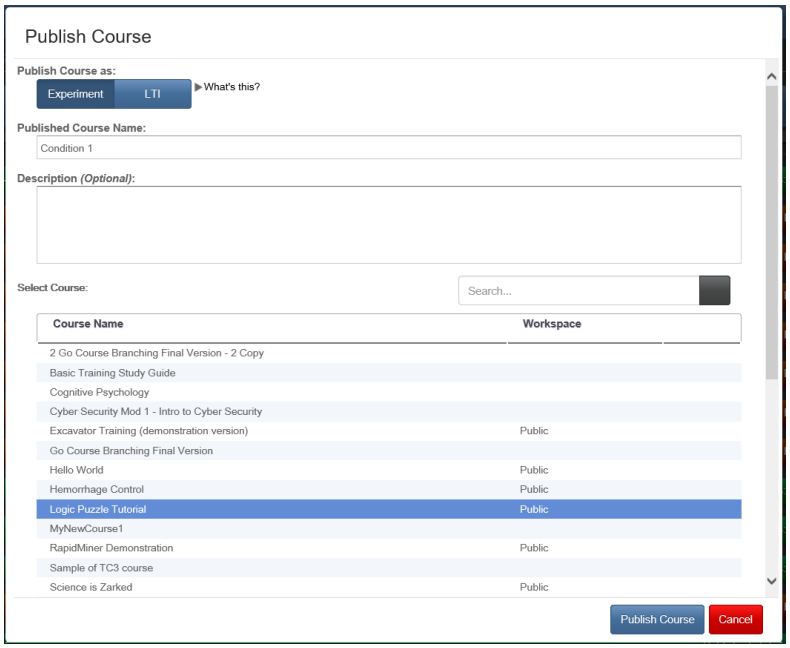


Figure 8. The Publish Course interface.

A URL will be provided that links students to that specific instance of the course so that they can interact with it. It is important to note that this publish functionality is copying the selected GIFT course at that moment in time. If any updates or edits are made to the original course file, a new published version of the course and URL must be produced in order for students to be able to interact with the new version.

After students have interacted with a course, an instructor can use the red and green course name interface in "Publish Courses" in order to extract data. Again, this design is primarily based on functionality that is of use in running experiments. See Figures 9 and 10 for an example of these interfaces. However, in order

for the data to be extracted, for the space instance, “Pause and Build Report” must be clicked and the data downloaded. To extract survey data, be sure to check the box that says “Survey Responses”, and in order to have each participant on one line in the exported document click “Merge each participant’s events into a single row”. This will then build a file that then needs to be saved to the computer. It will be a .CSV file which can then be opened in Excel. If the cells are merged, each student’s inputs will be listed on a single row. The tags that were added in the survey authoring process will be available at the top of each column. This is why it is very important to add tags to each of the questions, as otherwise numbers that do not provide information about the question’s content will be visible. Any needed post processing can then be completed on the file in Excel or the desired program.

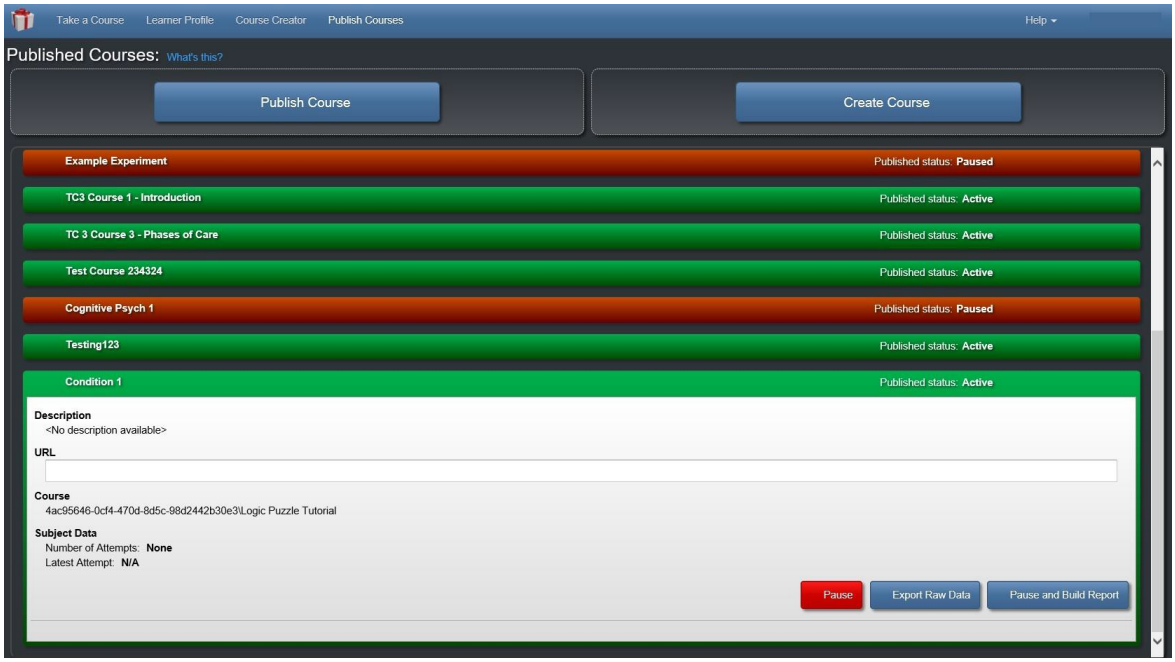


Figure 9. Published course interface

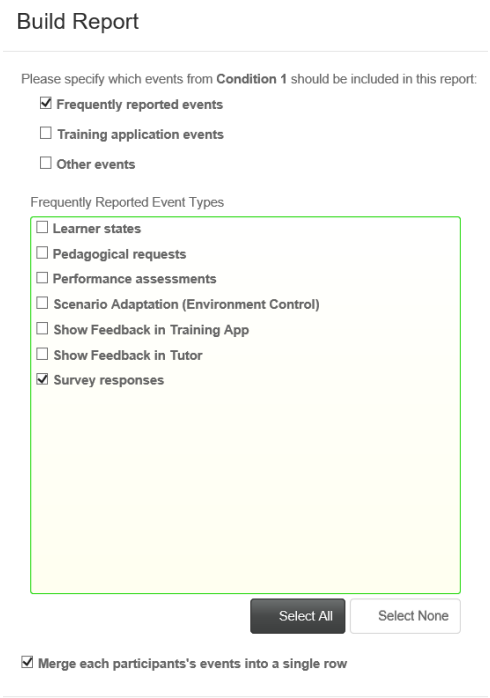


Figure 10. Build a report interface for extracting student data.

SUGGESTIONS FOR IMPROVEMENTS TO GIFT

As identified earlier in the paper, while there are many features of GIFT that are highly relevant for use by instructors, there are still some challenges to implementing it in an actual class.

Gradebook

Currently there is no easy way to have student learning outcomes populated into a course gradebook. It would be helpful to implement features into GIFT which provide this functionality for instructors. The most effective way to currently use GIFT for an assignment at current time would be as a pass/fail participation grade in which the instructor can do a quick export of the data and see that the student participated. If quiz scores need to be examined or calculated then it is more work for the instructor, and the format that is exported may not easily import into existing gradebooks that the instructor may be utilizing in other systems.

Student and Teacher Roles

In the current GIFT Cloud setup there is no way for the course creator (or even the student) to export data from the course that they have created while they are logged into their own account. It would be beneficial to have a function that is similar to that of the "Publish Courses" option which would allow for the specific data that has been generated from the course to be viewed from the main GIFT interface while logged in. One way to help implement this would be by creating teacher and student roles in the course. Then the course creator can indicate the GIFT account of the teacher, which could have access to the shared course's logs. GIFT has begun moving in this direction through the ability to share courses. The next step would be distinguishing between students and teachers and updating interfaces to reflect the role that the individual has in the system. This of course is not a straightforward task and will need to take into consideration special cases such as graduate students who are both instructors and students. Additionally, the implementation of a learning management system and potential gradebook features in GIFT could assist in addressing some of the current challenges to implementing GIFT in a class.

Test Bank Import

Many educational textbook publishers provide text banks which include questions that are associated with each chapter of content. These test banks often come in formats that are compatible with Learning Management Systems such as Blackboard and Webcourses. If these testbanks could be imported into GIFT it would save instructors time in inputting quiz questions. Additionally, there needs to be further clarity on the ability to reuse questions between surveys and question banks in GIFT in order to reduce mistakes when adding questions to the system. There is currently a Qualtrics import functionality for questions in GIFT. It may be helpful to identify textbook test bank formats and create import functions for them as well.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

While GIFT is a very powerful tool and can be of great use to instructors, there are still some challenges and function gaps that exist in GIFT. The implementation of a learner management system and persistent learner record store with the ability to provide information to instructors would be very helpful. Additionally, streamlining and clarifying the difference between Question Banks and Surveys would be helpful for instructors. The current guide provides recommendations, and instructions on how to implement GIFT in a class. The current optimal configuration to use would be to create a GIFT course which utilizes the Slide Show course object and distributing it to students using the "Publish Course" option. In order to facilitate using GIFT in a course, a few small feature updates could be made (e.g., gradebook/grade export, student/teacher roles) which would greatly improve the ease of use. GIFT is very useful in the current state, but with additional improvements it will become an even more powerful tool for instructors to utilize.

REFERENCES

Sinatra, A. M. (2014). The research psychologist's guide to GIFT. In Proceedings of the 2nd Annual GIFT Users Symposium (pp. 85-92).

Sinatra, A. M. (2015, August). The Instructor's Guide to GIFT: Recommendations for using GIFT In and Out of the Classroom. In Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym3) (p. 149).

Sinatra, A. M. (2016). The Updated Research Psychologist's Guide to GIFT. In Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym4) (p. 135).

Sinatra, A. M. (2018, May). The 2018 Research Psychologist's Guide to GIFT. In Proceedings of the Sixth Annual GIFT Users Symposium (Vol. 6, p. 259). US Army Research Laboratory.

Sottilare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). An updated concept for a Generalized Intelligent Framework for Tutoring (GIFT). GIFTtutoring. org.

ABOUT THE AUTHOR

Dr. Anne M. Sinatra *is part of the adaptive training research team within CCDC Soldier Center STTC's Learning in Intelligent Tutoring Environments (LITE) Lab, and works on the Generalized Intelligent Framework for Tutoring (GIFT) project. Her background is in Human Factors and Cognitive Psychology.*

ACKNOWLEDGEMENTS

The research described herein has been sponsored by the U.S Army Combat Capabilities Development Command – Soldier Center – Simulation and Training Technology Center. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

Enhancing GIFT Authoring User Experience through Interaction Design

Robert A. Sottilare, Ross Hoehn, Dar-Wei Chen, and Behrooz Mostafavi
Soar Technology, Inc.

INTRODUCTION

A major consideration in the design of tools that support authoring, instruction, deployment, and evaluation of adaptive instructional systems (AISs) is interaction design. According to Sottilare & Brawner (2018), AISs are artificially-intelligent, computer-based systems that guide learning experiences by tailoring instruction and recommendations based on the goals, needs, and preferences of each individual learner or team in the context of domain learning objectives. AISs include learning technologies that include intelligent tutoring systems (ITSs), intelligent mentors (recommender engines), and intelligent instructional me- dia. According to Preece, Rogers & Sharp (2007, p. 8), *interaction design* is defined as: "designing interactive products to support the way people communicate and interact in their everyday and working lives".

This paper specifically examines human interaction with processes enabled by the Generalized Intelligent Framework for Tutoring (GIFT; Sottilare, Brawner, Goldberg & Holden, 2012; Sottilare, Brawner, Sinatra & Johnston, 2017), an open-source architecture for authoring, deploying, autonomously managing, and evaluating adaptive instruction (e.g., ITSs that provide tailored instruction in a domain of knowledge – fundamental systems of the human body, rifle marksmanship or land navigation).

When we consider various aspects of AIS design, interaction design drives us to develop practices that will optimize user interactions during AIS authoring, deployment, instruction, and evaluation processes. We expect to develop recommended practices built upon a model of AIS users that considers:

- AIS user roles (e.g., learners, authors, instructional designers, system maintainers, and researchers)
- Capabilities and limitations of users in various roles
- Contributing factors to quality user experiences
- User feedback about their AIS experiences
- Measures of usability to compare/contrast alternative approaches

As it is for other systems, interaction design for AISs is a multidisciplinary process involving a variety of career fields (e.g., psychologists, computer programmers, and engineers), academic disciplines (e.g., human factors, cognitive psychology, social sciences, and informatics), and design practices (e.g., graphic design, conceptual modeling, engineering design, and product design). Next, we consider the influence of various disciplines (including computer science, psychology, instructional design, and cognitive science) in the AIS processes noted above with the goal of enhancing AIS user experiences through improved usability.

AIS USABILITY GOALS

Per Preece, Rogers, & Sharp (2007, p. 20), system usability is defined as "ensuring that interactive products are easy to learn, effective to use, and enjoyable from the user's perspective". Preece et al (2007) also associate the following goals/measures with usability that we have applied to GIFT and AISs:

- Goal: highly effective – a measure of how well an AIS is at doing what it was designed to do (e.g., improves knowledge and skills in a particular domain)
- Goal: highly efficient – a measure of how well an AIS is at supporting users in completing tasks or reaching goals
- Goal: high utility – a measure of the extent to which an AIS provides appropriate capabilities to meet user needs or desires
- Goal: easy to learn – a measure of how quickly a user can reach proficiency and use AIS capabilities
- Goal: easy to recall – a measure of how easy it is to remember AIS capabilities once learned

Now that we have identified usability goals associated with the interaction design of AISs, our next step is to apply these goals to the specific processes within GIFT and many AISs to identify interaction design gaps.

APPLYING AIS USABILITY GOALS TO GIFT PROCESSES

In this section, we begin to examine the interaction design of GIFT in terms of GIFT processes and their usability. While we understand that GIFT is a baseline concept or prototype, it does have a level of maturity (Technology Readiness Level 5 or 6 – Mankins, 1995) and a sufficient user base that warrants this examination. We also understand that while GIFT functions may not be present in all AISs, they are representative of AISs in that they have processes and common AIS functional components (i.e., learner model, instructional model, domain model and interface model). A goal of this section is to examine the usability of GIFT authoring tools, adaptive instruction, and evaluation tools through the lens of Nielsen & Molich's (March 1990) methodology for the heuristic evaluation of user interfaces.

Heuristic evaluation is usually an informal method of evaluating usability where a number of evaluators are presented with an interface design and asked to comment on its ease-of-use as it relates to a set of rules or criteria. In our case, we chose the following heuristics and provide feedback on usability with respect interaction design:

Simple and natural dialogue – GIFT does not provide any mechanism to insure simple and natural dialogue is authored. Feedback though a virtual human (VH) interface and text chat window is primarily at the discretion of the author. However, the VH interface could be greatly improved to engage the learner. Increasing the size of the VH in the dialogue window and the ability to swap out VH personas might be critical for interaction with learners during courses and experiments.

Speak the user's language/Be consistent – Given GIFT is a multi-disciplinary tool, its taxonomy is expected to be familiar to computer programmers, research psychologists, and instructional designers. This might be a bridge too far. We recommend GIFT adopt the ontology being developed under IEEE Project 2247. The AIS concept modeling subgroup is working with an interdisciplinary group of professionals to develop this ontology to support a common language for AIS authoring, deployment, automated instruction, and evaluation. For more information about the IEEE Project 2247, please visit: http://sites.ieee.org/sa- groups-2247-1/

Minimize user memory load – The authoring process in GIFT has been simplified over the last few years to provide a simple drag and drop interface, and the course objects have labels to identify their functions. This reduced working memory load during the authoring process. However, the optimal order for authoring is not specified and the state of authoring is not displayed for the user. We recommend an authoring dashboard to inform the user about any gaps in the development of a GIFT course or experiment.

Provide feedback – GIFT provides a tool to validate each course and provide textual feedback. We recommend a dashboard with a graphical indicator of authoring tasks and the percentage of the authoring tasks completed. A simplified graphical flow chart would also be useful to the author in tracking progress for the development of courses. GIFT also provides tools to validate and preview courses.

Provide clearly marked exits – In the GIFT course creator (authoring tools), there is a clearly labeled over-head menu with options to navigate from the course creator. Additionally, the upper right corner features a user profile dropdown with a logout command.

Provide shortcuts – No shortcuts are presently available for menu items. However, there are right-click options for edit, delete, and copy for all course objects in the course creator and for whole course in the "take a course" page. We recommend relabeling the "copy" option to read "copy and paste" for consistency.

Good error messages/Prevent errors – The GIFT authoring tools provide a tool to allow the author to preview their course from a specified starting point (e.g., beginning or adaptive courseflow object). Errors result when GIFT attempts to preview the course from a position where input is expected from the learner and was not provided. The preview simply aborts and does not provide an indication of the type of error experienced. An authoring dashboard that indicates common errors would also be useful.

In the next three subsections we examine specific GIFT functions with respect to usability and accessibility: authoring tools, courses, and evaluation tools.

Usability of GIFT Authoring Tools

Based on the large number of publicly available adaptive courses in the GIFT Cloud, the diversity of do- mains that those courses represent, the large number of experiments conducted using GIFT courses, and the drag-and-drop nature of the GIFT Authoring Tools (GAT), we characterize the GAT as having a high degree of usability. It is safe to say that the authoring tools enable users to construct relatively simple knowledge-based products without the need for programming skills, but that their usability with respect to more complex skill-based tasks leveraging external environments (e.g., serious games) are more difficult. Certainly the GIFT authoring tools have been used to construct some very complex tutors for land navigation and rifle marksmanship, but the construction of similar ITS is likely beyond the capabilities of most users.

The tools provide an easy-to-use method to sequence, configure, and modify course objects that represent a variety of content (e.g., media, assessments, surveys, and courseflow objects). A big asset is the survey authoring system which enhances user efficiency by allowing the import of surveys developed in Qualtrix. Another asset is the presence of a validation tool that highlights missing elements in the configuration of course objects. What is not readily apparent is the sequence of identifying learning objectives and linking them to content objects and learner states. Learning objectives, or in GIFT parlance – *concepts*, must be defined first and then media objects, assessment or survey questions, and courseflow objects can then be configured so they are associated with these concepts.

Three difficult authoring tasks in GIFT involve: 1) association of data sources with learner states, 2) association of content meta-data attributes with learner attributes and instructional phases, and 3) assessment of conditions in external environments.

Association of data sources with learner states – There is a need to define processes to acquire data (e.g., via sensors, surveys, self-reported data) to support classification of current learner states (e.g., engagement, arousal, motivation or domain knowledge) and/or predict future learner states.

Association of content meta-data attributes with learner attributes and instructional phases – There is a need to efficiently tag media and other content to aid content searches in the context of learner conditions and instructional

phases (e.g., Merrill's (1983) Rule Quadrant or Chi & Wylie's (2014) Interactive-Constructive-Active-Passive (ICAP) model).

Assessment of conditions in external environments – External environments (e.g., simulations such as Virtual Battle Space) are one method to provide interactive learning experiences. There is a need to develop methods to extract external environment conditions so AISs can support learner assessments without the need for computer programming; we recommend a utility to author condition classes automatically based on author specification.

Accessibility of GIFT Authoring Tools

Overall, GIFT and its affiliated authoring tools are extremely accessible in the cloud at https://cloud.gifttu- toring.org. As a cloud-based tool-suite, GIFT provides a scalable architecture that can grow easily with the number of users and the number of courses being developed. Developers can store their courses or experiments in the cloud and provide links to students or participants (Figure 1). Experimental data is stored in the cloud, can be configured for easy analysis using the Event Report Tool (ERT) and exported on demand for further analysis.

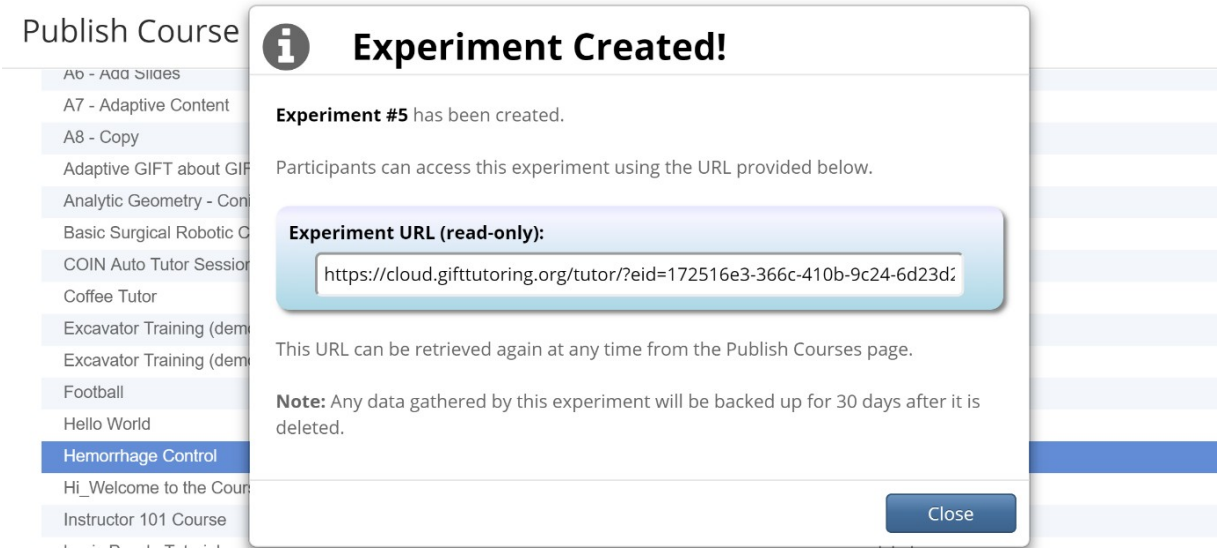


Figure 1. Cloud-based Courses and Experiments

Usability of Adaptive Instructional Courses in GIFT

After logging into the GIFT Cloud-based platform, each learner is introduced to their personalized GIFT Dashboard and a complete tiled list of course materials. They are presented with both title and thumbnail images which are accessible by the individual learner/author. Upon selection of a course, the installation of any necessary plug-ins (e.g., Java applet to run Microsoft PowerPoint) prompt the learner to run and install required software. This may be distracting, but for the most part can be avoided by using Microsoft PowerPoint slideshows that are converted to images.

GIFT is designed to provide a self-contained platform for AIS material authoring, curating and sharing. To satisfy this requirement, the cloud-based GIFT provides support for Java, UnityGL and embedding of API- based external media. There is great advantage to this capability in that GIFT can be used to point to existing media which satisfies legal access under Fair Use, "a doctrine in the law of the United States that permits limited use of copyrighted material without having to first acquire permission from the copyright holder" (Leval, 1989). Another advantage is that the author of the media retains control and provides updates. The GIFT ITS author is not required to manage the configuration of external media when the tutor simply points to the location of the media.

GIFT facilitates the presentation of instructional material to learners through embedded graphical features, such as slides and videos. Static material is presented in a fashion akin to PowerPoint materials. Individual learning objectives can comprise an entire course, a series of slides or a single slide. Video-based materials can be managed via API-embedding of videos hosted on YouTube. Video controls are native to GIFT, and allow the learner to control their receipt of the video presentation.

Exploiting the GIFT's use of UnityGL, course materials can be generated in and, thereby, delivered via an interactive synthetic instructor. These instructors present course particulars to the learner by an instructor- avatar reciting scripted materials. Synthetic instructors are also capable of presenting prescript dialogue- based options to the learner via multiple-option input, permitting the learner to direct their own exposure to course materials. Also empowered by UnityGL, a course can contain simulations (*e.g.* a training simulation for operating an excavator) capable of being entirely contained within GIFT.

Figure 2. Virtual Human interface in GIFT courses
Left = current interface; Right = recommended more engaging interface

Usability of Evaluation Tools in GIFT

After a course has been developed using the GIFT authoring tools, the course can be published as a experiment or self-contained tutoring platform. Experiments can be hosted on the GIFT Cloud, which facilitates both the collection format and the storage of data. These experiments will appear in the "Publish Courses" tab for management and data retrieval. Selection of an active experiment will provide a link to the course materials to be given to learners, subject metrics (e.g., latest attempt, number of attempts), and a course description. This experiment management environment assists in course sharing, validation, error checking, and metadata tagging.

Within this management screen, collected experimental data can be exported for analysis and publication. Exportation of raw data will generate a copy of a JSON database for all metrics collected while learners were interacting with the course. This database provides extensive information that can be employed in defining and deriving bespoke metrics not yet available in GIFT. A researcher can also have GIFT automatically generate reports containing predefined metrics.

IMPROVING THE INTERACTION DESIGN OF GIFT

In this section, we provide five major recommendations for improving the interaction design of GIFT authoring processes for instruction and experimentation.

Recommendation #1 – Author Dashboard

The usability of the GIFT authoring tools could be greatly improved by the addition of a checklist or author dashboard that highlights prerequisite events in the form of a nodal network (Figure 2).

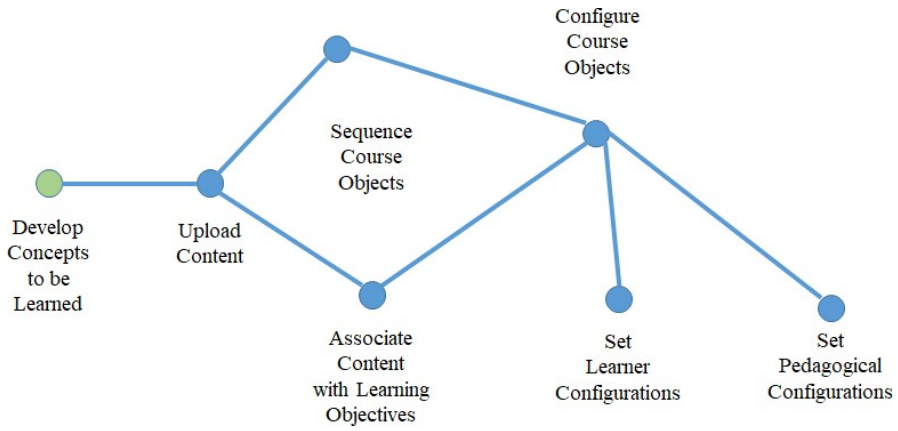


Figure 3. Notional GIFT Author Dashboard Widget: Nodal network depicting authoring tasks and progress

Recommendation #2 – Collaborative Authoring

The usability of the GIFT authoring tools could be greatly improved by expanding the authoring interface for use by multiple authors/instructional designers/content developers.

Recommendation #3 – Researcher Dashboard

The usability of the GIFT ERT used to configure experimental data could be greatly improved through a researcher dashboard that allows researchers to use a WYSIWYG interface to configure data for analysis and reporting.

Recommendation #4 – Condition Class Dashboard

A condition class dashboard (Figure 4) would automate the creation of condition classes to support the acquisition of measures and the assessment of learners with respect to learning objectives.

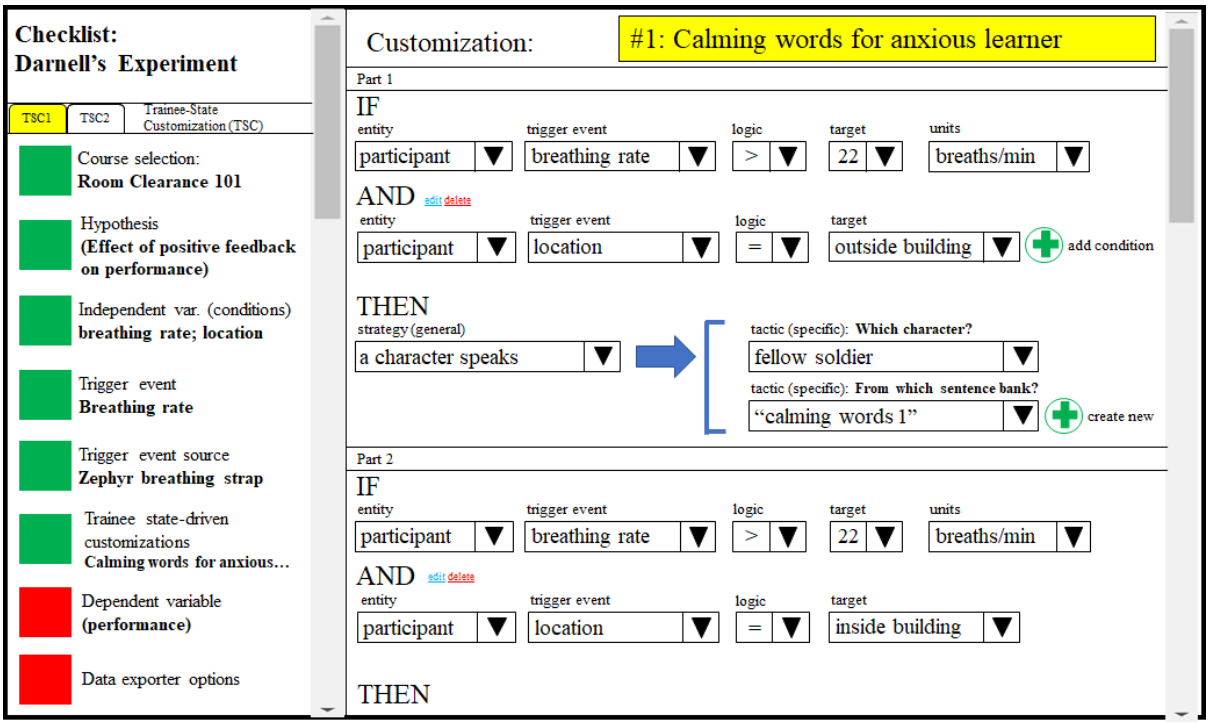


Figure 4. Notional GIFT Condition Class Dashboard: Provides ability to identify data sources and automate building of condition classes

The communication of data moving between AISs (such as GIFT-based ITSs) and external environments (e.g., serious games, virtual simulations, cases or problem databases) are difficult to facilitate manually, but may lend themselves to a repeatable, automated process. Currently, GIFT provides a mechanism called *condition classes*, which are specific statements that allow a program to check a condition and execute certain parts of code depending on whether the condition is true or false. Condition classes contain the instructions for how GIFT should respond to data from external environments, including strategies or tactics that occur in the environment (e.g., increasing the number of pedestrians in a city block when the learner's performance in a surveillance task moves from moderate to high). Today, a computer programmer must manually generate these condition classes. We are suggesting that researchers' abilities to create training materials and experiments might be enhanced by allowing the author to structure conditions in an intuitive dashboard and then automatically generate the condition classes needed, all without knowledge of programming.

This dashboard should implement easily-understood non-programming features (e.g., questions written in prose, if-then statements, drop-down menus of current GIFT-compatible equipment, auto-complete, GUI elements, wizard-style interactions) that a researcher can use to fully input information related to their de- sired condition class. Then, when the researcher has provided all of the relevant information for the condition class, the dashboard should connect to a condition class generator that outputs the JavaScript code for the conditions defined and saves it for other researchers to use in the future. This example solution would expedite training-related research and widen the potential user base of GIFT.

Recommendation #5 – Population Model Dashboard

A population model dashboard would allow authors to import historical data or build population profiles relevant to the knowledge, skills, abilities, attitudes and other characteristics (KSAAOs) of a group or population. Population models are useful in enabling standards, identifying trends, evaluating learner with respect to their peers, and making instructional decisions based on norms. Population models may also be useful in bootstrapping instructional decisions when specific individual learner data is unavailable. We advocate that mechanisms to support/generate instructional decisions be examined within the larger field and context of distributed learning. As such, authoring tools should be data-driven and utilize information from a wider set of domain-independent learning resources. Managing such a wide set of information re- quires constant analysis of instructional and media factors, methods of assessment across a range of environments, and comparisons of required and actual KSAAOs across a variety of domains in order to build population models for structuring pedagogical policies.

Population models are statistical distributions of various learner and team experiences and achievements observed within a population. Populations can be simple or complex, singular or nested within each other. Relating these models within and between domains can facilitate the generation and definition of concepts, assessments, and instructional decisions within the GIFT authoring tools. Population models may include hierarchical cluster distributions of KSAAOs to identify learning gaps and develop recommendations for future experiences, identification of concept dependencies and pre-requisite relationships to support the formation of concept maps for sequencing learning experiences, gap analysis summaries between individual learners and their peer groups at various echelons, and analysis of existing population models with respect to individual and group outcomes to increase the accuracy and precision of the statistical distributions.

CLOSING THOUGHTS

To a great extent, the ease of use for the GIFT authoring tools could be greatly improved and expanded to support the development of more complex tutors. Through the use of visualization (e.g., dashboards) and the development of methods to automate steps in the authoring process, we might realize additional efficiencies. Dashboards, scripts, and, machine learning techniques like genetic algorithms could enhance the authors' efficiency through guided authoring, automated processes, and improved situational awareness. We highly recommend research to develop tools and methods in support of the authoring process to expand the usability and applicability of GIFT to a larger set of educational and training domains.

REFERENCES

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4), 219-243.

Leval, P. N. (1989). Toward a fair use standard. *Harv. L. Rev.*, *103*, 1105. Mankins, J. C. (1995). Technology readiness levels. *White Paper, April*, *6*, 1995.

Merrill, M. D. (1983). Component display theory. *Instructional-design theories and models: An overview of their current status*, 1, 282-333.

Nielsen, J., & Molich, R. (1990, March). Heuristic evaluation of user interfaces. In Proceedings of the SIGCHI con-ference on Human factors in computing systems (pp. 249-256). ACM.

Preece, J., Rogers, Y., & Sharp, H. (2007). *Interaction design: beyond human-computer interaction*. John Wiley & Sons.

Salas, E. (2015). Team training essentials: A research-based guide. London: Routledge.

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., Holden, H. K. (2012). The Generalized Intelligent Framework for Tutoring (GIFT). Concept paper released as part of GIFT software documentation. *U.S. Army Research Laboratory— Human Research & Engineering Directorate (ARL-HRED)*, Orlando, FL, USA.

Sottilare, R., Brawner, K., Sinatra, A., Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Frame- work for Tutoring (GIFT). *US Army Research Laboratory*, Orlando, FL, USA.

Sottilare, R., Brawner, K. (2018, June). Component Interaction within the Generalized Intelligent Framework for Tutoring (GIFT) as a Model for Adaptive Instructional System Standards. In the Adaptive Instructional System (AIS) Standards Workshop of the *14th International Intelligent Tutoring Systems (ITS) Conference*, Montreal, Quebec, Canada.

ABOUT THE AUTHORS

Dr. Robert Sottilare *came to SoarTech as the Science Director for Intelligent Training in 2018 after completing a 35- year career in federal service in both the Army and the Navy training science and technology organizations. Most recently, he led adaptive training research at the US Army Research Laboratory where the focus of his research was automated authoring, instructional management, and analysis tools and methods for intelligent tutoring systems (ITSs). He is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT), an open source, AI-based adaptive instructional architecture used for learning sciences research. GIFT has over 2000 users in 76 countries. He has over 200 technical publications in the learning sciences field (over 1500 citations in the last 5 years). His doctorate is in Modeling & Simulation with a focus in Intelligent Systems from the University of Central Florida.*

Dr. Ross D. Hoehn *is a research scientist in SoarTech's Intelligent Training division. He earned is Ph.D. in Theoretical Chemistry from Purdue University in 2014, and continued research in chemical physics, quantum information and artificial intelligence until 2018. His research areas include: adaptive artificial intelligence, generative AI techniques, team learning and training, pedogogy, biological-based agent simulations, swarm mechanics, quantum information science, quantum computing and quantum mechanically-driven biophysical phenomenon. He was an active researcher and manager of the NSF Center for Chemical Innovation: Quantum Information for Quantum Chemistry, a multi-million dollar multi-year research effort centered at Purdue University to utilize quantum computing for quantum mechanical calculations.*

Dr. Dar-Wei Chen *is a research scientist in SoarTech's Intelligent Training division where he specializes in applying human factors principles to learning environments. He previously earned a doctoral degree in engineering psychology from the Georgia Institute of Technology and his research led him to be named a finalist for the James D. Foley Scholarship (top Ph.D. students in design and technology at Georgia Tech), recipient of the Larry S. O'Hara Scholarship (top senior Ph.D. student in the College of Sciences), and a recipient of the Georgia Tech Presidential Fellow- ship. His experience with GIFT includes a summer stint with the Learning in Intelligent Tutoring Environments (LITE) Laboratory at the Army Research Laboratory where he used a GIFT-powered simulated shooting range to train cadets on fundamentals of marksmanship.*

Dr. Behrooz Mostafavi *is a research scientist in SoarTech's Intelligent Training division. He earned his Ph.D. in Computer Science from North Carolina State University in 2016, with a concentration in intelligent and adaptive instructional systems. He continued research in applying personalized learning techniques to a growing tutoring framework for discrete mathematics at the undergraduate level during his postdoctoral work at NCSU's Center for Educational Informatics before joining SoarTech in 2018. His experience with GIFT includes recent in-depth analy- sis and experimental design of mid-level reporting techniques, as well as a study of GIFT and related technologies, techniques, and applications, and written contributions to the Army Research Laboratory Design Recommendations for Intelligent Tutoring Systems series of books.*

Extending GIFT Wrap to Live Training

Fleet C. Davis[1], Jennifer M. Riley[2], and Benjamin S. Goldberg[3]
BMT Inc.[1], Design Interactive Inc.[2], CCDC-Soldier Center, STTC[3]

INTRODUCTION

The Generalized Intelligent Framework for Tutoring (GIFT) is a modular suite of capabilities aimed at overcoming the challenges associated with authoring and delivering computer-based instruction via an intelligent tutoring system (Sottilare, Brawner, Goldberg, & Holden, 2012). One of the primary objectives for GIFT development is to create an integrated, user-friendly authoring experience that is training platform agnostic. Humanproof, with teammate Design Interactive, recently completed the fourth generation GIFT Wrap prototype, a software application that allows training developers to configure the real-time, automated delivery of instructional content triggered by assessing state changes within the training application's learning environment and/or learner. The following sections summarize previous GIFT Wrap development efforts including the first three generations that focused on developing authoring tools across virtual- and game-based training environments, provide an overview of the fourth generation of GIFT Wrap for extending the functionality to live training environments, and discuss future applications of GIFT Wrap.

BACKGROUND

Evolution of GIFT Wrap Development

The GIFT Wrap project was a multi-year, research and development effort aimed at developing a fully-integrated, user-friendly tool for authoring individual, adaptive training following a Crawl-Walk-Run (CWR) approach to training (Goldberg, Davis, Riley, & Boyce, 2017) that would employ multiple training applications. For the purposes of scoping this project, Army Map Reading and Land Navigation training (Department of the Army [DA], 2007) was selected as the primary use case.

The first and second generation of GIFT Wrap laid the foundation for the iterative development of the tool. These first two generations were aimed at overcoming the challenges associated with authoring a Domain Knowledge File (DKF) (Shute, Ventura, Small & Goldberg, 2013) and the disconnect between GIFT authoring tools and training application content creation tools (Davis, Riley, & Goldberg, 2017). The resulting proof-of-concept provided a "blended authoring environment" that allowed users to author real-time assessments directly within the context of a training application's virtual environment (i.e., the Augmented Reality Sandtable (ARES) terrain map) (Hoffman, Markuck, & Goldberg, 2016) and a completely redesigned user interface (UI) for authoring a Domain Knowledge Files (DKF) (Davis et al., 2017) (see Figure 1). At this stage, GIFT Wrap supported the Crawl phase of skill acquisition focused on the fundamentals of Map Reading and Land Navigation.

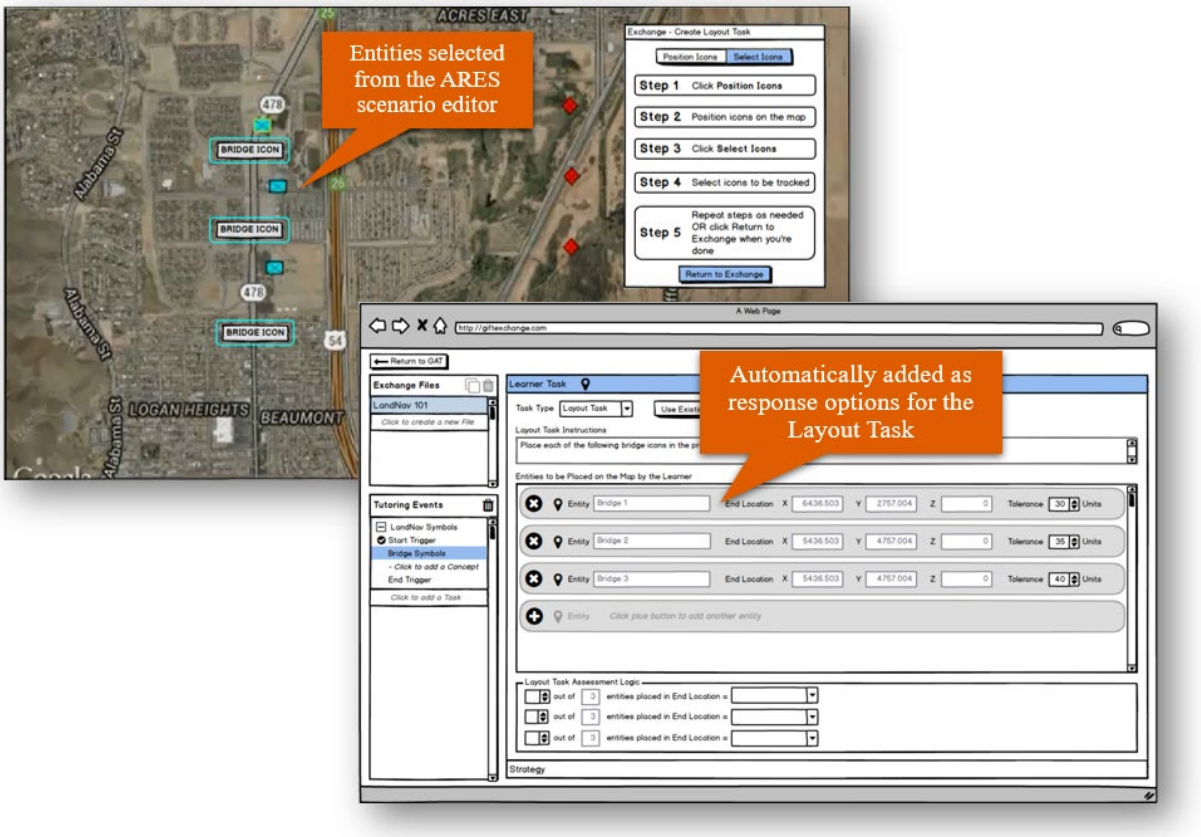


Figure 1. Authoring in ARES – 2nd Generation GIFT Wrap

The third generation of GIFT Wrap incorporated additional features for authoring DKFs and extended the blended authoring experience to include the LandNavHD Unity game, a computer-based land navigation trainer used as a practice environment for dead reckoning procedures (Davis, Riley, & Goldberg, 2018) (see Figure 2). The third generation of GIFT Wrap supported the Walk phase of skill acquisition focused on applying Map Reading and Land Navigation knowledge within an interactive exercise.

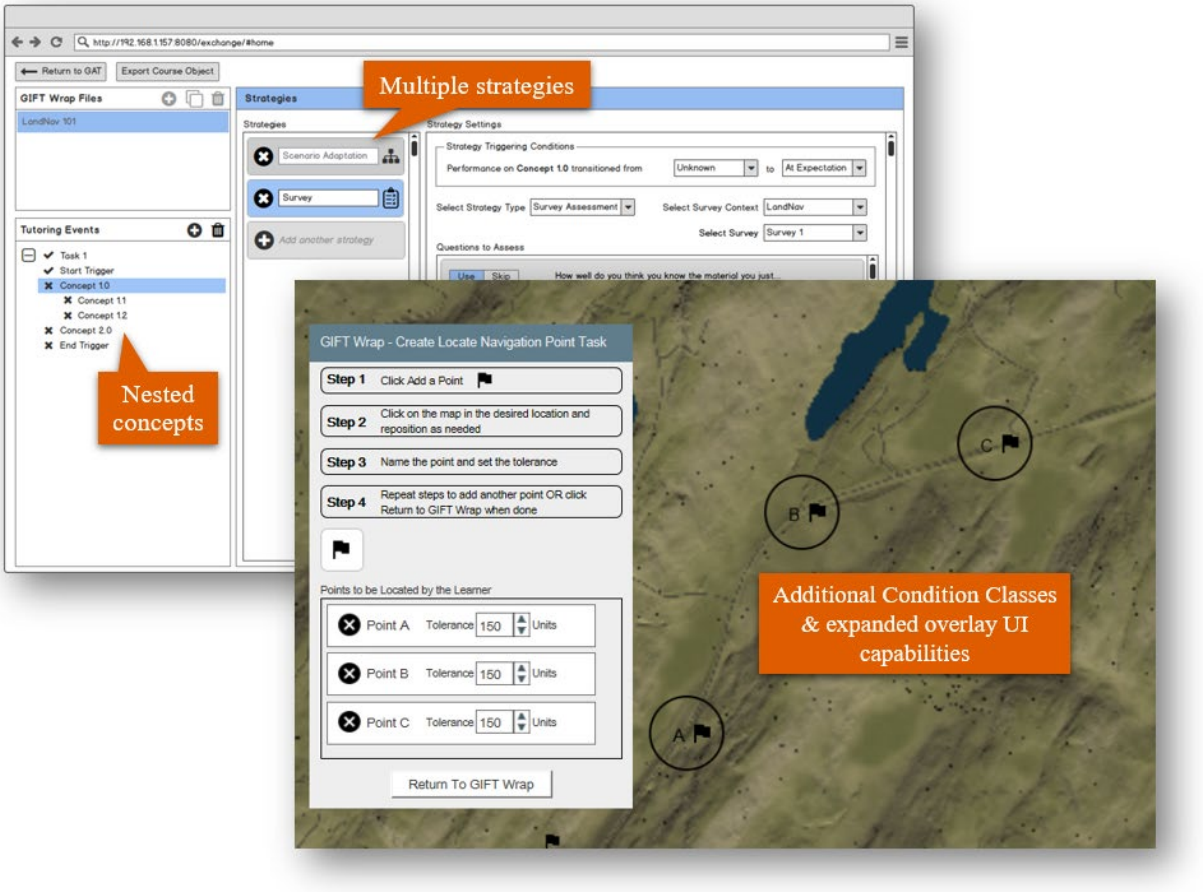


Figure 2. Authoring in LandNavHD – 3rd Generation GIFT Wrap

GIFT Wrap within the Course Development Process

GIFT course development is primarily supported through the GIFT Course Creator. With this interface, the author sets the course properties, defines the course objects, and applies relevant media to the instructional elements and approaches. The Course Creator provides interface components to set variable parameters for the provided course objects. GIFT Wrap provides additional interfaces to set parameters that are specific to the implementation of real-time assessments that occur during the completion of the overall course.

A few elements should be accounted for in the application of real-time assessments that relate directly to utilization of GIFT Wrap in the scenario development process. The author should have defined the course "concepts" to be addressed and the course objects that will have real-time assessments associated with them. Based on the defined concepts and corresponding assessments, the relevant training application(s) should be selected. Two approaches are presented here as examples on GIFT Wrap applications.

First, the author can include a course object referencing the direct application of a training applications (e.g., Virtual Battlespace (VBS), ARES, PowerPoint) for presenting a real-time assessment to the learner. In the same manner, the author adds the relevant course object for a training application within the course flow. The course object presents selections, for example, to identify specific scenarios to be used within the training application and the real-time assessment to be applied. GIFT Wrap would be invoked for setting up the relevant parameters of the real-time assessment – the concept to be assessed, the manner of assessment, the criterion for triggering the assessment, the criterion for assessing performance, and the strategy to be applied given the performance outcomes. The user can set

up the parameters prior to using GIFT, or with the direct integration of GIFT Wrap with the GIFT software, the author can access GIFT Wrap when setting up the real-time assessment with the training application course objects.

A second application for using GIFT Wrap is the application of an "adaptive course flow" object. This is a course object that facilitates adaptive behaviors within the course, including providing remedial instructional content for concepts for which the learners fail to reach a pre-set performance criterion. As with the real-time assessment, the course concepts must be defined. With this course object, the author can elect to push the learner through a "practice phase". This practice phase is essentially a strategy that is presented if performance on the "check on learning" element of the adaptive course flow does not reach desired levels. The "practice phase" initiates practice within a training application. As with the real-time assessment setup, when the training application is applied, the author would utilized GIFT Wrap to define the practice, which is essentially the tasks to be completed to rehearse and assess the skill associated with the concept, as well as to define the criterion for whether or not the desired performance is achieved with the practice. If the performance is acceptable, per the parameters set by the author, the learner is allowed to move ahead in the course. If the performance is not acceptable, the learner remains within the flow or loop of the adaptive course flow object.

In either case of the course setup, GIFT Wrap is used when the author is setting up parameters associated with activities in a selected training application. If the author has pre-defined the concepts, the tasks of interest, and the other elements of interest to a practice phase or real-time assessment (e.g., areas of interest, points or locations of interest, entities of interest), GIFT Wrap can be used to define those before the development of the GIFT course. If the author has not pre-defined each element of the course, they can use GIFT Wrap as the course is being set up, moving between the GIFT Wrap and the Course Creator as needed to set up parameters for real-time assessments and adaptive course flow practice opportunities.

FOURTH GENERATION GIFT WRAP

Extending the Blended Authoring Experience to Support Live Training Exercises

The GIFT Wrap blended authoring experience was extended beyond the ARES and LandNavHD training applications to live training exercises authored within Google Maps. Using the Google Web Toolkit (GWT), the GIFT Wrap Overlay UI was integrated with Google Maps allowing users to author real-time assessments on any area of the map (see Figure 3).

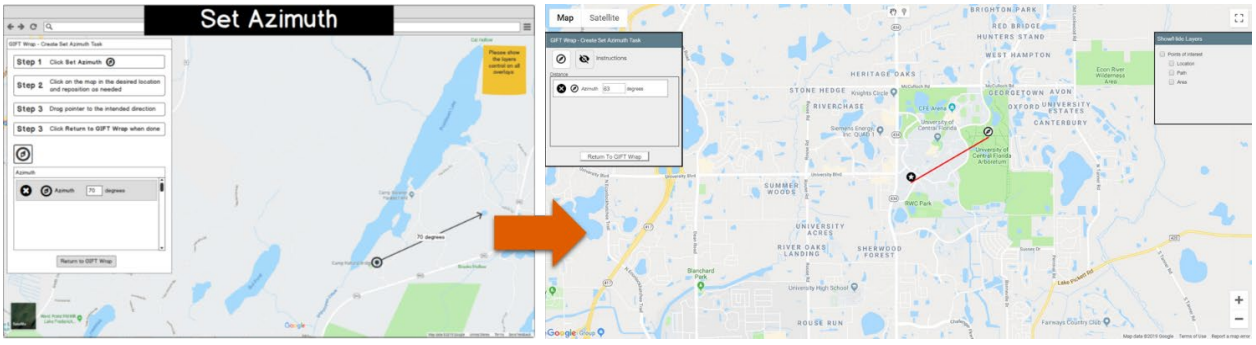


Figure 3. Example of Overlay UI Design Translated to Functional Implementation

In addition to GIFT Wrap integration with Google Maps, new real-time assessments were created to mimic the instructional approaches used during "Terrain Walk" exercises at the United States Military Academy at West Point. These new real-time assessments include:

- **Grid Coordinate** – Assesses the accuracy of the grid coordinate entered by a Learner for a given location.
- **Identify Terrain Features** – Assesses the Learner's ability to identify the location of terrain features on a map.
- **Orient Map** – Assesses the Learner's ability to orient a map correctly using the direction the Learner's mobile device is pointing as a proxy for the direction they are facing with the map.
- **Pace Check** – Assesses the Learner's ability to measure a straight-line distance using the pace count method.
- **Predict Distance** – Assesses the Learner's ability to determine the distance between two given points.
- **Set Azimuth** – Assesses the Learner's ability to set a given azimuth on their compass. This assessment uses the direction the Learner's mobile device is pointing to determine the Learner's bearing or azimuth.

The GIFT Wrap main page UI and Overlay UI were modified to include authoring for each these new assessments. Figure 4 show an example of authoring the "Identify Terrain Features" assessment using the GIFT Wrap UI integrated with Google Maps.

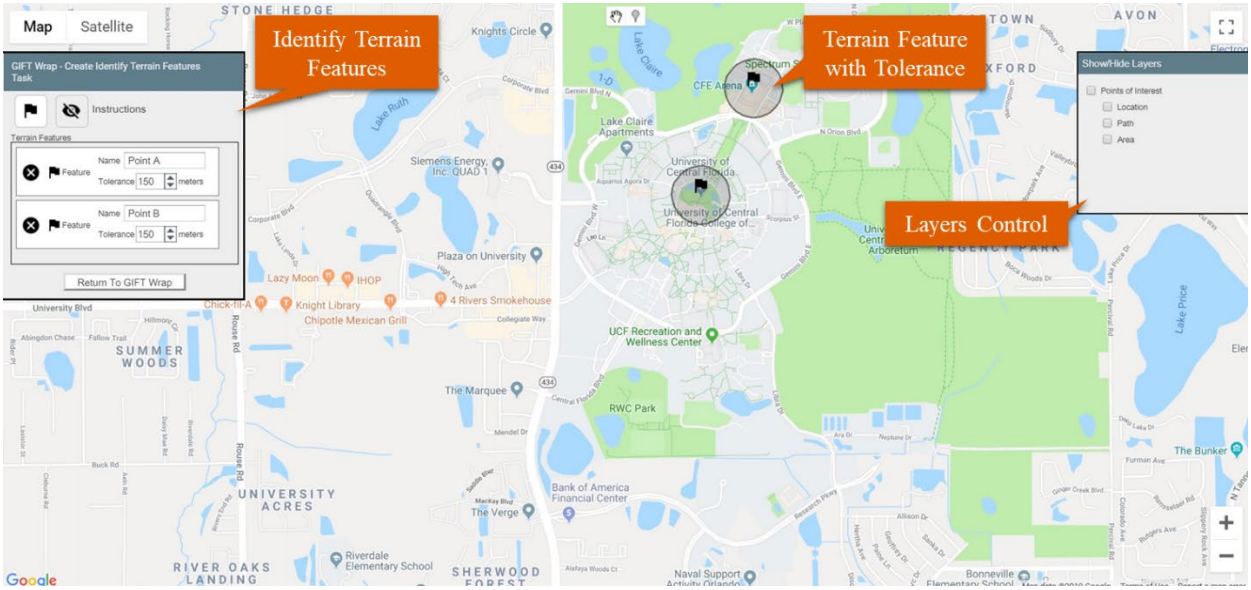


Figure 4. Example Overlay UI for Identify Terrain Features Real-time Assessment

These assessments were intended to be representative of typical training activities on a navigation course in a real-world environment (i.e., the Run phase of skill acquisition). However, instead of the training being delivered by an instructor, the trainee would experience the adaptive training provided by the GIFT Tutor User Interface (TUI) delivering the course via mobile device, such as smartphone (see Figure 5).

Figure 5. Notional Mock-up of the Mobile TUI

Other GIFT Wrap Enhancements

In an effort to continuously improve the usability of GIFT Wrap, the following features were added to the fourth-generation design:

- **Show/hide Instructions** – This feature allows GIFT users to view instructions when needed or to reduce the occlusion of the map by the Overlay UI by closing instructions when not needed. This is a particularly useful feature for use with the Google Maps application as it does not support a "movable" overlay, but requires the overlay be docked in place.
- **Layers** – Considering that users might want to see multiple elements on the same map to visualize the "Terrain Walk" activity, the design was expanded to include controls to show/hide "Layers" which coincide with three common land navigation assessment elements that may be references for multiple condition classes – points, path, and areas.
- **Points of Interest** – This feature allows users to add common map features or elements that may be reused and/or referenced by multiple real-time assessments. Multiple points, paths, and/or areas can be set and later selected for use for the set-up of real-time assessments.

Integration with Google Maps

The GIFT Wrap integration with Google Maps required some workarounds due to incompatibilities. Because Google Maps utilizes a JavaScript application programming interface (API) for web development and GIFT utilizes GWT, software engineers were required to use a third-party tool that allow us to integrate Google Maps with GWT. This API (GWT-MapsV3-Api, 2019) essentially creates a bridge between Google Map's JavaScript API and the GWT framework, thus allowing the GWT developer to implement Google Maps. Once the bridge was in place, the features that could be leveraged for GIFT Wrap were identified by evaluating the relevant land navigation tasks. Based on the real-time assessment to be supported, several tools were selected for implementation. The work then turned to determining how to display the relevant Google Map controls in the GIFT Wrap overlay control panel. The overlay control panel houses the instructions and controls that facilitate the authoring of "learner tasks". For real-time assessments, the overlay captures the parameters authored for the task and links the parameters of the task with the Google Maps API in order to pass the desired performance data captured during training to GIFT.

The GWT solution provided a proof-of-concept for the approach, though some limitations are evident. For example, Google Maps offers JavaScript API for web, not GWT. The GWT library provided fewer options and an older library which did not allow for use of the most recent Google Maps API features.

Integration with GIFT Baseline

The merge included integrating the technical capabilities previously developed for utilizing LandNavHD and Google Maps for real-time performance assessment. Details on the structure and format of new conditions classes were provided. The GIFT development team provided feedback on expanding Unity applications to support future overlay authoring, so the software was modified to make future expansion easier.

A few challenges were identified. First, a goal for GIFT is to limit the number of conditions classes through re-use across training applications. The applicability and implementation of the condition classes however are influenced by the context and specifics of the domain. Also, for conditions classes referencing points of interest and other map- and position-based elements, the coordinate system(s) in place can impact capacity to re-use without some modification or specification for the selected coordinate system. The newly defined conditions classes are under review to determine if they can be directly re-used.

Versioning and dating issues cause some confusion in merging with the baseline code. Changes in features and functions of variable GIFT versions required repeated modifications to GIFT Wrap code to maintain compatibility. Though the modifications are published by the GIFT development team, it is difficult to know which aspects of the GIFT Wrap are affected until tests were completed or inconsistencies in functioning were observed. This is likely to be an issue for others developing on a separate GIFT code branch and then executing a merge.

Lastly, there were challenges to overcome with implementation and testing GIFT wrap with the mobile version of GIFT and the newly defined condition classes. To deploy GIFT to mobile device, the team used the "Publish Course" feature that generate the Uniform Resource Locator (URL) for the course. From the URL, the course can be accessed. GIFT Cloud allows this feature to be accessed anywhere as long as the user has access to the link. Initially, the version the team developed produced a local URL, which meant that only the local owner's computer could access the course. The resolution was to launch to Amazon cloud in order to get access the published course URL, just as with the cloud version of GIFT, so that we could test implementation on the mobile phone.

LIMITATIONS AND CHALLENGES

As the GIFT Wrap design has evolved, each iteration of the tool has added features and functionality that incrementally reduced the burden associated with authoring real-time assessments and configuring the delivery of instructional strategies across several training applications. However, there are still many technical challenges to overcome. The following sections describe the current limitations of the tool and some of the future development challenges.

GIFT Wrap Interoperability Limitations and Integration Challenges

As previously described in this paper, GIFT Wrap integration with third-party systems (e.g., Google Maps) remains challenging. While the creation of the Gateway Module and various plugins has allowed for interoperability and reduced development time, there is still a considerable amount of customization required to establish the communication between GIFT Wrap and a training application that is necessary for implementing real-time performance assessments. For example, the third generation of GIFT Wrap was integrated with the LandNavHD training application. The real-time assessments users could author for this training application (e.g., Avoid Area, Locate Navigation Points) required positional data from LandNavHD that were not included in the existing GIFT

Unity plugin. Several new event handlers had to be added to the plugin that sent messages to GIFT providing information used for real-time assessment.

There is also the challenge of integrating GIFT Wrap authoring capabilities with those of the training application. One of the goals for the GIFT Wrap project was to create a "blended authoring environment" that would allow users to author real-time assessments within the context of a training application's content creation tools via an Overlay UI (Davis et al., 2017). The intent was to merge the GIFT Wrap UI with the content creation tool's UI in such a way that users would perceive the tool as one, seamless authoring experience. However, there are two significant challenges to implementing this design. First, some training applications simply lack scenario authoring capabilities. In these cases, workarounds are required to implement the authoring UI as intended. For example, in the case of the LandNavHD Unity game, a top-down image of the terrain was extracted, and a new layer was created in the GIFT Wrap UI to simulate the functionality of authoring within the training application's virtual environment. Second, for training applications that do including content creations tools, it's likely that access to the source code is needed in order to integrate GIFT Wrap functionality. For example, VBS is used by the Army for land navigation training and includes content creation tools (e.g., the Offline Mission Editor (OME)) for creating and editing the scenarios. The GIFT Wrap Overlay UI could potentially be integrated with the VBS OME such that the user could reference elements of the VBS scenario (e.g., waypoints, navigation flags) for real-time assessments (e.g., Locate Navigation Points). However, without access to the proprietary VBS source code, it is impossible to implement this functionality and create a seamless authoring experience for the user.

Authoring Limitations and Challenges

Maximizing usability has been a major focus throughout the development of the GIFT Wrap design. As such, several usability evaluations were conducted (Davis et al., 2018) as GIFT Wrap gradually incorporated DKF Authoring Tool (DAT) functionality and added new authoring capabilities. However, as computer-based tutoring system (CBTS) capabilities continue to advance and the intended use of GIFT Wrap broadens, some of the existing UI features may not be able to accommodate these new use cases. Real-time assessments are becoming more robust, training scenarios are growing in complexity, and the potential applications of CBTS are expanding beyond individuals and small teams to multi-echelon, collective training in new domains. Authoring tools such as GIFT Wrap will need to be modified and, in some cases, completely redesigned to accommodate these changes without compromising the usability of the tool.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

Building on the first three generations of GIFT Wrap, the fourth generation was aimed at extending GIFT's authoring capabilities for live, land navigation training via integration with Google Maps and GIFT mobile capabilities. With the completion of this integration and new authoring functionality, GIFT Wrap is now capable of supporting a CWR approach to training for the Map Reading and Land Navigation use case.

The most recent GIFT Wrap development efforts have focused primarily on authoring individual, adaptive training for Map Reading and Land Navigation. Many of the existing GIFT Wrap authoring capabilities and corresponding Condition Classes are applicable and/or easily modifiable to accommodate new training applications. Future research should concentrate on extending GIFT Wrap beyond the current use case to other Army elements (e.g., squad, platoon) for collective training and to other training applications across Army domains (e.g., armored, mission command). For example, GIFT Wrap location-based assessments could be slightly modified and used to author real-time assessments for a company team practicing a wedge formation and/or adjusting to the appropriate formation under different circumstances.

Future research should also be done to continuously examine and iteratively improve the GIFT Wrap user experience as use cases continue to be added. For example, the GIFT Wrap UI could be modified to facilitate authoring and visualizing dependencies amongst numerous events including triggering events, team behaviors and/or performance,

the impact of instructional strategies, etc. Alternative UI designs should be considered to ensure that the GIFT Wrap tool is flexible and robust enough to accommodate future applications including operational requirements for the Army's Synthetic Training Environment (STE) capability.

REFERENCES

Davis, F., Riley, J.M., & & Goldberg, B. (2018). Iterative Development of the GIFT Wrap Authoring Tool. In Proceedings of the Sixth Annual GIFT Users Symposium (GIFTSym6).

Davis, F., Riley, J.M., & & Goldberg, B. (2017). Development of an Integrated, User-Friendly Authoring Tool for Intelligent Tutoring Systems. In Proceedings of the Fifth Annual GIFT Users Symposium (GIFTSym5).

Department of the Army (2007). Map reading and land navigation (FM 3-25.26). Washington, DC.

Goldberg, B., Davis, F., Riley, J. M., & Boyce, M. W. (2017, July). Adaptive training across simulations in support of a crawl-walk-run model of interaction. In *International Conference on Augmented Cognition* (pp. 116-130). Springer, Cham.

GWT-MapsV3-Api (2019). Retrieved from https://github.com/branflake2267/GWT-Maps-V3-Api

Hoffman, M., Markuck, C., & Goldberg, B. (2016). Using GIFT Wrap to Author Domain Assessment Models with Native Training Applications. In Proceedings of the Fourth Annual GIFT Users Symposium (GIFTSym4).

Shute, V., Ventura, M., Small, M., & Goldberg, B. (2013). Modeling Student Competencies in Video Games Using Stealth Assessment. In R. Sottilare, A. Graesser, X. Hu & H. Holden (Eds.), *Design Recommendations for Intelligent Tutoring Systems, Volume 1: Learner Modeling* (pp. 141-152).

Sottilare, R.A., Brawner, K.W., Goldberg, B.S. & Holden, H.K. (2012). The Generalized Intelligent Framework for Tutoring (GIFT). Concept paper released as part of GIFT software documentation. Orlando, FL: U.S. Army Research Laboratory – Human Research & Engineering Directorate (ARL-HRED). Retrieved from: https://gifttutoring.org/attachments/152/GIFTDescription_0.pdf

ABOUT THE AUTHORS

Mr. Fleet Davis *is a Senior Human Factors Engineer at BMT, Inc. He served as the Principal Investigator for the GIFT Wrap project.*

Dr. Jennifer Riley *is the Performance Augmentation Division Head at Design Interactive, Inc. She served as the Co-Principal investigator for the GIFT Wrap project.*

Dr. Benjamin Goldberg *is an adaptive training scientist at the Army Research Laboratory's SFC Paul Ray Smith Simulation & Training Technology Center. He leads research focused on instructional management within ARL's Learning in Intelligent Tutoring Environments (LITE) Lab and is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT).*

GIFT as a Framework for Self-Improvable Digital Resources in SIAIS

Xiangen Hu[1,2], Zhiqiang Cai[1], Arthur C. Graesser[1], and Jody L. Cockroft[1]
The University of Memphis[1], Central China Normal University[2]

INTRODUCTION

There are four components in a minimalist model of Self-Improvable Adaptive Systems (SIAIS) (Hu, Tong, Cai, Cockroft, & Kim, 2019). These four components are 1) human learners that are assumed constantly self-improving (i.e., learning); 2) self-improvable learning resources that are either human resources (trainers/teachers, for example) or digital resources such as digital tutors which are capable of changing (improving) constantly; 3) learning environments that are diverse physical or virtual locations, contexts, and cultures in which students learn; 4) learning processes that are instructional sequence for any given domain for particular learner groups (such as grades). One publicly available important framework for self-improvable digital resource in SIAIS is Generalized Intelligent Framework for Tutoring (GIFT), which is an empirically-based, service-oriented framework of tools, methods and standards to make it easier to author computer-based tutoring systems (CBTS), manage instruction and assess the effect of CBTS, components and methodologies ("Overview - GIFT - GIFT Portal," n.d.). In this paper, we take AutoTutor as an example of self-improvable tutoring systems and discuss the rule GIFT may play as a framework of self-improvable learning resources.

SELF-IMPROVABLE LEARNING RESOURCES

Based on Hu et al (in press), *self-improvable learning resources* are defined as those learning resources that can update, retrieve, and utilize their associated memory of the learning activities. The human learner is obviously self-improvable and is constantly working to self-improve. Human teachers/trainers, and human study mates are also self-improvable and constantly improving as the result of constant interacting with human learners. Unfortunately, not all digital learning resources are self-improving. Relevant to the focus of the current paper, we are interested in specially designed digital resources that are self-improvable. As an example, AutoTutor is one such specially designed digital resources.

AutoTutor

AutoTutor (Graesser, Wiemer-Hastings, Wiemer-Hastings, & Kreuz, 1999; Nye, Graesser, & Hu, 2014) is an intelligent tutoring system that holds conversations with the human learner in natural language. AutoTutor has produced learning gains across multiple domains (e.g., computer literacy, physics, critical thinking). Three main research areas are central to AutoTutor: human-inspired tutoring strategies, pedagogical agents, and technology that supports natural language tutoring ("AutoTutor," n.d.). For the purpose of the current paper, we list a few less known properties of AutoTutor that make it a self-improvable digital resource in SIAIS. These properties make AutoTutor a variable controlled system. To illustrate, we list a few variable controllable components of AutoTutor that can influence the behavior of AutoTutor:

Answer Grading Model. AutoTutor conversation is often referred to as Expectation-Misconception Tailored (EMT) conversation, in which a human learner learns by constructing an acceptable answer to a main question through answering a sequence of hint/prompt questions asked by AutoTutor. An AutoTutor main question usually requires an answer of 3 to 10 sentences. A hint/prompt question targets one aspect of the answer. The answer to a hint question is usually a sentence or a clause, while an answer to a prompt is usually a word or phrase. AutoTutor intelligently selects hint/prompt questions based on the learner's input, which could be a good answer, a partial answer, a misconception, an irrelevant answer or even not an answer (e.g., a question). The AutoTutor answer grading model

is responsible for classifying the learner's inputs. The model is trained through deep learning neural network (Cheng, Cai & Graesser, 2018) with semantic features. For each newly developed AutoTutor application, AutoTutor uses a pre-trained model for early answer grading. When enough learning data is collected, the domain specific model is trained. The model could be further improved when more data is collected. Thus, the answer grading model is considered as an important variable that can be changed over time to improve AutoTutor's performance.

Avatars: AutoTutor employs several conversational avatars when interact with students. The avatars can play different roles such as computer tutors or computer students. Each of the avatars can have an assigned "personality" serving different functions during tutoring. For example, a tutor avator could have warm, neutral or cold personality; a peer student avatar may have the personality as a student leader, a hard working learner, an aggressive competitor, etc. The "personality" of an avatar is reflected by its facial emotions and its commonly used speeches, called "canned expressions". The avatar face can be selected from available avatar library. The canned expressions can be revised over time. Thus, avatars are a variable in AutoTutor.

Scripts: AutoTutor uses author prepared scripts for avatars. However, when learning data is collected, the scripts could be changed over time. For example, useless speeches could be removed; inadequate questions and speeches could be revised; and missing questions and speeches could be added. Moreover, typical utterances from human learners could be added as speeches for avatars that play the role of peer students.

Rules: AutoTutor uses a set of "if-then" rules to determine what to do next in any given state. A state is determined by the learning history (what has happened so far) and the current input, including natural language input and "world events", such as an interaction between the learner and an interactive element on the application interface, a time controlled change in the learning environment, etc. A rule set is often embedded with pedagogical strategies. For example, a vicarious learning rule set supports conversation between a tutor avatar and a peer student avatar, with minimal involvement of the human learner. A tutoring rule set specifies the way how to interact with learners who have medium level knowledge about the topic under discussion. A teachable agent rule set provides learners the opportunity to learn through teaching a peer student avatar. Each rule in a rule set can be changed over time. The criteria used to select a rule set is also changeable. Thus AutoTutor conversation rules are also a variable.

GIFT as a framework for Self-Improvable digital resources in SIAIS

There are many learning resources like AutoTutor that can be integrated into GIFT. GIFT framework requires any ITS based on GIFT (GIFTITS) is an integration of four core modules ("Overview - GIFT - GIFT Portal," n.d.): **Sensor Module, Learner Module**, **Pedagogical Module**, and the **Domain Module**. The current prototype of GIFT (https://cloud.gifttutoring.org/) is a GIFTITS. The Sensor Module has interfaces to support commercial sensors (e.g., Affectiva Q-Sensor) and its function is to format, process and store sensor data. The Domain Module provides domain content to support training, assesses trainee performance against standards, and provides domain-specific feedback to the trainee when the Pedagogical Module identifies the need for feedback based on trainee performance. The Trainee Module uses trainee performance, historical data (e.g., past performance) and sensor data to determine the trainee's cognitive and affective state. Current implementation of GIFT is primarily for content authoring and resource integration. Each of these Modules is interchangeable through the virtue of interfacing standards. This allows each Module designer to select the type of approach that they believe is suited towards instruction. For instance, a sample configuration may have a webcam sensor that interprets Facial Action Units (FACs), a rule-based performance assessment, a Feedback Generation Engine that generates varying levels of hints upon request, a finite state machine of trainee assessment, and pedagogy that gives hints on failed problems.

Relevant to the focus of the current paper, one very important implementational properties of the current GIFT prototype is its modularity. All modules of GIFTITS are variable controlled. A each module is controlled by an XML file. For example, a domain knowledge file (DKF) contains the information needed to execute on a single lesson. Learner Configuration File is an XML that configures the learner module to support building learner states from inputs such as sensor data and performance assessments. There are configuration files for sensor module (SensorConfigurationFile), pedagogy module (PedagogicalConfigurationFile) that are controls behavior of GIFTITS

when interact with learner. In addition, other variables are also separately specified (in common.properties file). This implementation properties of the GIFT prototype shows that GIFTITS can be variable controlled hence self-improvable.

SELF-IMPROVABILITY OF LEARNING RESOURCES

It was pointed out earlier that *self-improvable* learning resources are defined as those learning resources that can update, retrieve, and utilize their associated memory of the learning activities (Hu et al., 2019). Having variable controlled components will only make a digital resources *self-improvable* but not necessarily self-improving. Other key properties are needed. In the case of AutoTutor, its self-improvability is due to three key factors: a) AutoTutor is a cloud-based implementation with constant connection with a Learner Record Store (LRS) (Nye et al., 2014). b) Behaviors of AutoTutor are variable controlled. c) The variables that control AutoTutor's behaviors could be changed based on the behavioral data collected in LRS. The same is for GIFT. Current implementation of AutoTutor only has a). The self-improvability of AutoTutor depending on b) and c). Only proper implementation of b) and c) can make AutoTutor self-improving.

Self-improvability of GIFTITS

We have argued that ITS implementation based on GIFT (as shown in the current GIFT prototype implementation) may have all modules and components variable controlled hence self-improvable. To make GIFTITS truly self-improving, additional key properties need to be added:

1. System behavior of GIFTITS and human learner interaction behavior should be captured and stored within the same data scheme (such as xAPI). Current used of the behavioral data are collected mostly for post-hoc analysis. When the data is used to make GIFTITS self-improvable, there are special requirements, For example, the speed of retrieving and processing data should be fast enough for real-time feed back to the GIFTITS. Because the data will be used to improve GIFTITS, additional requirement for the data schema need to be considered (Hu et al impress).

2. A collection of APIs need to be created that connect all variables of GIFTITS to the data store. These APIs will need to be constantly computing values based system behavior data and capable of real-time updating GIFTITS. The output of these APIs can either be an updated XML file (such as the DKF, PedagogicalConfigurationFile, or parameter values in the common.properties file).

With 1) and 2) can only make GIFTITS self-changable. There is no mechanism to guarantee the GIFTITS is actually improving the learning experience and effectiveness. So it is very important to ensure self-improvability is achieved. In order to make this happen, a set of theory-driven empirically verified ideal tutoring behaviors need to be specified parametrically. For example, based on Graesser et al. (2008) GIFTITS needs to ask deep questions to during tutoring session, so for effective ITS, there might be a minimum requirement for number of deep questions asked during a given period of time. In addition an effective ITS in a given domain may have an optimal combinations of questions at different levels (Graesser & Person, 1994). So in addition to 1) and 2) listed above, self-improvability of GIFTITS needs to have

3. A pre-set of ideal (effective and efficient) tutoring strategies specified computationally so it can be used to guide APIs of 2).

RECOMMENDATION AND FUTURE RESEARCH

Any GIFTITS can be self-improvable learning resource due to its design with variable controlled modules and components. Self-improving GIFTITS is possible if the self-improbability requirements (1-3) are met. Consider

building self-improving GIFTITS as ultimate goal, it is necessary to enhance GIFT framework with the three self-improbability requirements. Specifically,

1. A extended behavior xAPI-like data profile (Hu et al. in press) need to be created that is capable of capture all interactions between GIFTITS and human learner such that all system behavior of GIFTITS are captured similar to that of human learner's behavior.

2. A collection of optimum domain-specific task-dependent tutoring strategies need to be created. These optimum tutoring strategies are computationally specifiable. For example, if a conversation-based GIFTITS is created based on Expectation-misconception tailored (EMT) dialog (Olney, Graesser, & Person, 2010), there exists an optimum combination of hints, prompts, pumps, and elaborations (Graesser et al., 1999; Olney et al., 2010).

3. A set of APIs needed to be created. These APIs that are constantly monitoring GIFTITS behaviors and make real-time changes of variable values in GIFTITS modules and components of based on 2).

REFERENCES

AutoTutor. (n.d.). Retrieved April 13, 2019, from http://ace.autotutor.org/IISAutotutor/index.html

Graesser, A. C., Halpern, D. F., & Hakel, M. (2008). 25 principles of learning. Task Force on Lifelong Learning at Work and at Home Washington, DC.

Graesser, A. C., & Person, N. K. (1994). Question Asking during Tutoring. *American Educational Research Journal*, *31*(1), 104.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., & Kreuz, R. (1999). AutoTutor: A simulation of a human tutor. *Cognitive Systems Research*, *1*(1), 35–51.

Hu, X., Tong, R., Cai, Z., Cockroft, J. L., & Kim, J. W. (2019). Self-Improvable Adaptive Instructional Systems (SIAIS) -- A proposed model. In A.M. Sinatra, A. Graesser, X. Hu, V. Rus, A. Olney (Ed.), *Design Recommendations for Intelligent Tutoring Systems: Volume 7 -- Self-Improving Systems*. Orlando, FL: U.S. Army Research Laboratory.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education*, *24*(4), 427–469.

Olney, A. M., Graesser, A. C., & Person, N. K. (2010). Tutorial Dialog in Natural Language. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in Intelligent Tutoring Systems* (pp. 181–206). Berlin, Heidelberg: Springer Berlin Heidelberg.

Overview - GIFT - GIFT Portal. (n.d.). Retrieved February 21, 2019, from https://gifttutoring.org/projects/gift/wiki/Overview

ABOUT THE AUTHORS

Dr. Xiangen Hu *is a professor in the Department of Psychology, Department of Electrical and Computer Engineering and Computer Science Department at The University of Memphis (UofM) and senior researcher at the Institute for Intelligent Systems (IIS) at the UofM and is professor and Dean of the School of Psychology at Central China Normal University (CCNU). Dr. Hu received his MS in applied mathematics from Huazhong University of Science and Technology, MA in social sciences and Ph.D. in Cognitive Sciences from the University of California, Irvine. Dr. Hu is the Director of Advanced Distributed Learning (ADL) Partnership Laboratory at the UofM, and is a senior researcher in the Chinese Ministry of Education's Key Laboratory of Adolescent Cyberpsychology and Behavior.*

Zhiqiang Cai *is a Research Assistant Professor with the Institute for Intelligent Systems at the University of Memphis. He has a M.S. degree in mathematics received in 1985 from Huazhong University of Science and Technology, P. R. China. After 15 years of teaching mathematics in colleges, he has worked in the field of natural language processing and intelligent systems. He is the chief software designer and developer of Coh-Metrix, OperationAries, CSAL AutoTutor and many other text analysis tools and conversational tutoring systems. He has co-authored over 70 publications.*

Arthur C. Graesser *is a professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis and is a Senior Research Fellow in the Department of Education at the University of Oxford. He received his Ph.D. in psychology from the University of California at San Diego. Dr. Graesser's primary research interests are in cognitive science, discourse processing, and the learning sciences. More specific interests include knowledge representation, question asking and answering, tutoring, text comprehension, inference generation, conversation, reading, education, memory, emotions, computational linguistics, artificial intelligence, human-computer interaction, and learning technologies with animated conversational agents. He has published over 500 articles in journals, books, and conference proceedings.*

Jody L. Cockroft.*is a Research Specialist at the University of Memphis in the Institute for Intelligent Systems. Prior to joining the University of Memphis, she was with the University of Tennessee Health Science Center in Memphis where she was involved in with various clinical trials and bench research for over twenty years. She earned her A.A. from the University of Tampa and her B.S. from the University of Memphis. She has been working with the UofM team for the past five years on the Army Research Laboratory project on the Generalized Intelligent Framework for Tutoring (GIFT) and the Advanced Distributed Learning (ADL) Academy projects and the Advanced Learning Theories, Technologies, Applications and Impacts (ALTTAI) Consortium efforts. She is serving as the Treasurer for the IEEE Project 2247 Adaptive Instructional Standards working Group.*

THEME II:
AI, MACHINE LEARNING AND GIFT

Application Of Reinforcement Learning For Automated Contents Validation Towards Self-Improving Online Courseware

Noboru Matsuda and Machi Shimmei
Center for Educational Informatics Department of Computer Science
North Carolina State University

INTRODUCTION

Online education has been growing rapidly for the last decade with exponential growth of diverse students population (Shapiro et al., 2017) and adaptive technology enhancements (Lerís, Sein-Echaluce, Hernández, & Bueno, 2017). However, building practical online courseware is extremely costly—it requires extensive knowledge and expertise in theories of learning and teaching (Clark & Mayer, 2003; Slavich & Zimbardo, 2012). Most of the time, instructional designers and instructors design an initial courseware from their honest intuition, and then the courseware will be iteratively modified to meet better learning outcome. Though, iterative software engineering is a norm for almost any sort of practical software applications (Fishman, Marx, Blumenfeld, Krajcik, & Soloway, 2004), it requires significant knowledge to identify issues to be fixed for improvement.

It is therefore critical to develop a transformative theory of practical learning-engineering methods for iterative online courseware creation. Without such methods, it is not likely to have sustainable system of online education. What if the courseware improves itself over time?

The larger goal of our current project is to develop a self-improving online courseware that automatically detects and fixes ineffective parts of the existing courseware relative to students' learning achievement. As a step towards achieving this pivotal goal, we propose to develop an integrated development environment (IDE) where human and AI collaboratively build online courseware through iterative design engineering—a machine detects issues and a human fixes them. As a step towards the proposed human- AI collaboration, this paper describes an innovative application of a reinforcement learning technique called RAFINE (**R**einforcement learning **A**pplication **F**or **IN**cremental courseware **E**ngineering). The RAFINE aims to identify ineffective instructional elements on existing online courseware given a record of individual students' learning activity logs.

In the rest of the paper, we first discuss related works followed by a detailed description of RAFINE. We then describe details about a simulation study and results as a proof of concept.

RELATED WORKS

Reinforcement learning (RL) has been used for educational applications in particular to compute effective pedagogical strategies for adaptive tutoring. Previous works applied RL to find optimal pedagogical decisions such as teaching actions (Rafferty, Brunskill, Griffiths, & Shafto, 2015), hint messages (Martin & Arroyo, 2004), dialogue moves (Min Chi, VanLehn, Litman, & Jordan, 2011; Tetreault, Bohus, & Litman, 2007), learning activities (Shen & Chi, 2016), and navigation (Iglesias, Martinez, Aler, & Fernandez, 2009). Other studies have applied RL to compute effective domain models such as model solutions (Barnes & Stamper, 2008). The effects of educational RL policy have been tested both with real and simulated data where some studies showed a positive effect of the policy (Beck, Woolf, & Beal, 2000; M. Chi, Koedinger, Gordon, Jordan, & VanLehn, 2011) while others did not (Iglesias et al., 2009).

It is fairly common that the computed policies in the previous educational applications were optimized for learning outcome and learning time (Beck et al., 2000). To the best of our knowledge, the previous works are all mostly about computing the optimal pedagogical decisions. No research has been conducted to apply RL to identify ineffective instructional elements. Furthermore, under the framework of the ordinal RL, the rejected instructional contents do

not necessarily have a flaw—the second best might be as effective as the best. The current paper demonstrates *how RL can be applied to identify ineffective instructional contents on existing online courseware.*

SOLUTION: RAFINE

Overview of the RAFINE Method

We consider the RAFINE method as Human-AI collaboration to improve the quality of existing online courseware. An initial version of the online courseware will be used by students and their activities will be logged. These activity data consist of standard clickstream data and students' responses (and their correctness) for formative assessments. Since students' activity data show a chronological record of their behavior on the online courseware, we call them the *learning trajectory* data hereafter.

The RAFINE method first consolidates learning trajectories collected from *all* students into a single state transition graph, called a *learning trajectory graph* (LTG), and annotates the states with predefined rewards. A value iteration technique is then applied to compute a *converse policy* that shows the worst activities to be taken to achieve the expected learning outcomes. As a consequence, the converse policy corresponds to a set of instructional elements that have the least likelihood to contribute to students' learning.

Our central hypothesis is that those instructional elements that *frequently* appear as a converse policy across different states in a given LTG are likely to be ineffective and hence the subject for refinement. Those instructional elements identified as ineffective will then be presented to courseware developers as a recommendation for a courseware modification. The RAFINE method will be iteratively applied to the revised courseware by collecting a new batch of learning trajectory data to further improve the courseware.

Model Representation

The unit of analysis of the RAFINE method is an *instructional element* that constitutes online courseware. In the current study, we deal with three types of instructional elements: (1) videos, (2) formative assessments (aka quizzes), and (3) hint messages associated with formative assessments. We assume that all assessment quizzes are equipped with hint messages.

Let Φ be a set of instructional elements appearing in the given learning trajectories. We assume that the target courseware was used by a large number of students hence Φ contains all instructional elements on the target online courseware. Let a_i^T, a *learning activity*, be an instructional element taken (e.g., watching a video or answering a quiz) by student i at time T. Let LT_i be a *learning trajectory* for student i who has n_i learning activities. LT_i is a chronological record of learning activities:

$$LT_i = \{a_i^1, \ldots, a_i^{n_i} \mid a_i^t \in \Phi, \quad t = 1, \ldots, n_i\}.$$

We assume a presence of a *skill model* that contains a set of skills each representing a unit of knowledge that students have to learn, aka knowledge components (Koedinger, Corbett, & Perfetti, 2012). This assumption implies that each instructional element is tagged with a single skill in the given skill model. The RAFINE method is applied to each individual skill separately. Let Φ^μ be a set of instructional elements for skill μ. Learning trajectories are also broken down into individual skills. Let LT_i^μ be the learning trajectory that contains only learning activities about skill μ. The RAFINE method must be applied to each bundle of Φ^μ and LT_i^μ for all μ separately. A single application of the RAFINE method identifies ineffective instructional elements relative to a particular skill.

For a sake of simplicity without a loss of generality, let's assume that there is only one skill in our target online courseware. We therefore eliminate the skill index from Φ and *LT* in the following descriptions unless otherwise desired for a clarification.

In the learning trajectory graph, states represent *learning status* and edges represent learning activities taken that caused a change in status. We define a *learning status* for student *i* at time *T* for a particular skill μ as an intermediate state of learning represented as a pair of Action History and Mastery Level;

$<\boldsymbol{ah}_{i,T}, p_{i,T}(\mu)>$. Action History $\boldsymbol{ah}_{i,T}$ is a binary vector $<ah_i^1, \ldots, ah_i^K>$ where ah_i^m shows whether student *i* has taken the *m*-th instructional element in Φ^{μ} by time *T* (assuming the instructional elements are ordered and $|\Phi^{\mu}| = K$).

Mastery Level $p_{i,T}(\mu)$ is a scalar value showing a predicted probability of student *i* applying skill μ correctly, should he/she answer an assessment quiz for the skill μ at time *T*. The value of Mastery Level is rounded down to the nearest multiple of 0.05 (e.g., 0.18 becomes 0.15). Mastery Level, $p_{i,T}(\mu)$, will be computed based on the history of learning activities with an underlying assumption that commitment to a learning activity for a particular skill would increase Mastery Level by a specific amount. There are several known techniques available to achieve this goal including Bayesian models (e.g., Corbett & Anderson, 1995) and regression models (e.g., M. Chi et al., 2011). As long as Mastery Level is monotonically updated, any student-modeling technique would work for the RAFINE method.

To consolidate individual students' learning trajectories into a single learning trajectory graph (LTG), each individual student's learning trajectories are first converted into a *learning trajectory path*. This is done by chronologically traversing a learning trajectory while creating states each representing an intermediate learning status $<\boldsymbol{ah}_{i,T}, p_{i,T}(\mu)>$. While traversing the learning trajectory, $\boldsymbol{ah}_{i,T}$ and $p_{i,T}(\mu)$ are updated accordingly. For example, assume there are six instructional elements: Video1, Video2, Quiz1, Quiz2, Hint1, and Hint2. A state *s* <101000, 0.40> indicates that a student had watched Video1 and took Quiz1 before reaching the state *s*. It also indicates that a predicted Mastery Level at the time of arriving at the state *s* was 0.4. Assume that the student answered Quiz1 incorrectly to reach the state *s*. Now, the student needed to review Hint1, which caused a transition from *s* to *s*' where *s*' is <101010, 0.45> with an assumption that reviewing a hint increased the Master Level by 0.05.

All individual students' learning transition paths are then aggregated into an LTG by merging the same states. As a consequence, the states in the LTG generally have multiple incoming and outgoing edges. Note that in the LTG, student and time (i.e., the parameters *i* and *T* in an individual student's learning trajectory path) are abstracted. Therefore, in the following explanations, a tuple representing a state is denoted as $<\boldsymbol{ah}, p(\mu)>$. In an LTG, the states where the value of the Mastery Level, $p(\mu)$, is greater than a pre-defined threshold (which is usually 0.85) are called *terminal states*—meaning that students became proficient in applying skill μ. All outgoing edges at terminal states are discarded.

Rewards

A reward value of a particular state depends on the Mastery Level, $p(\mu)$, both at the current and successor states. As an example, consider two students who landed on the same state *s*, but then took different learning activities. One student reached a successor state by answering an assessment quiz incorrectly (i.e., $p(\mu)$ was not increased) whereas the other student watched a video (i.e., $p(\mu)$ was increased).

In our model, a reward for state *s* where the student took a learning activity *a* to reach a successor state *s*' is defined as:

$$(s,a,s') = \begin{array}{ll} -0.14 & (m(s) = ml(s^{\mathrm{A}}) < 0.85) \\ -0.05 & (ml(s) < ml(s') < 0.85) \\ 0.95 & \mathrm{F}0.85 \leq m(s^{\mathrm{A}})\mathrm{H} \end{array}$$

In the equations above, *ml*(*s*) returns the Mastery Level at the state *s*. A reward at the state *s* becomes the greatest when the successor state is a terminal state. Otherwise, the rewards are set to be small negative values so that the RL would find the shortest path to a terminal state while computing a policy as shown in the next section. We assume that the Mastery Level grows monotonic, i.e., students never unlearn. Therefore, a reward where *ml*(*s*) > *ml*(*s*') is undefined.

Converse Policy

Given the reward function *R* as mentioned above, a value function for state *s* under a policy π is defined as follows, where S is a set of all states in a given LTG:

$$V^{\mathrm{K}}(s) = \sum_{\mathrm{PA}\in\mathrm{Q}} T(s, \pi(s), s')(\mathrm{R}(\mathrm{s}, \pi(s), \mathrm{s}^{\mathrm{A}}) + \gamma V^{\mathrm{K}}(s'))$$

In the current implementation, the discount factor γ is arbitrarily set to be 0.9. A transition model *T*(*s*, *a*, *s*') is derived from the learning trajectory data collected from actual students as the probability of students reaching state s' when they took a learning activity a at state s.

In general, a policy suggests an action to be taken in a certain state to maximize the value function (Wiering & van Otterlo, 2012). However, for the purpose of Rafine, we need to know which instructional elements should not be taken—i.e., we need to know which action has the least expected reward. There- fore, through the value iteration, the value function is updated as follows where *A*(*s*) shows a set of actions appearing in outgoing edges at state *s*:

$$(s) \leftarrow \min_{\mathrm{V}\in\mathrm{W(P)}} \sum_{\mathrm{PA}\in\mathrm{Q}} T(s, a, s')(R(s, a, s^{\mathrm{A}}) + \gamma V(s'))$$

After the value function is converged, the action that minimizes the value function for state *s* is identified. We shall call this policy the *converse policy*:

$$(s) = \operatorname*{argmin}_{\mathrm{V}\in\mathrm{W(P)}} \sum_{\mathrm{P}\setminus\in\mathrm{Q}} T(s, a, s')(R(s, a, s^{\mathrm{A}}) + \gamma V^{\mathrm{K}}(s^{\mathrm{A}}))$$

EVALUATION STUDY

Our central hypothesis is that those instructional elements that *frequently* appear as a converse policy across different states in a given LTG are likely to be ineffective and hence should be revised. To test this hypothesis, we conducted an evaluation study with hypothetical learning trajectories generated by simulated students.

Although any instructional element can be selected as a converse policy, the current version of RAFINE only includes videos and hints in its recommendation. This is because there are known quantitative methods, e.g. item response theory (Baker, 2001), that can be used to evaluate the quality of assessment items.

Three instances of online courseware were created to control the quality of courseware with varying ratios of a number of effective instructional elements to all instructional elements on the courseware. We assumed that there was only one skill involved in the mock online courseware. All three instances of courseware had the same structure: they consisted of three pages (Page0, 1, 2), and each page included three lecture videos and three formative assessments (i.e., quizzes). All quizzes had hints associated. All instructional elements on the mock courseware (9 videos and 9 hints total) except assessment quizzes (for the reason mentioned above) were coded as either effective or ineffective. The high-quality courseware had a 8:1 split (8 effective video / hint and 1 ineffective video / hint);

the moderate-quality courseware had a 4:5 split; and the low-quality courseware had a 1:8 split. Let's call them H (High), M (Moderate), and L (Low) courseware hereafter.

Simulated students started from Page0 and randomly took a total of 10 to 14 instructional elements. At least two instructional elements must be taken to proceed a page. When simulated students answered a quiz incorrectly, they were forced to review the associated hint and take the same quiz again. The simu- lated student's performance on the assessment quizzes was determined by their latent proficiency that indicates a probability of answering a quiz correctly. In the real world, the latent proficiency increases according to the actual learning activities taken and student's latent trait of learning that determines the learning rate. To simulate the growth in the latent proficiency $p_{i,T}(\mu)$, we used a logistic regression model representing a probability of student i answering a quiz about the skill μ correctly at time T as shown below:

$$p_{\#,\wedge}(\mu) = \_\frac{1}{1 + e^{\mathrm{ab}_{\text{„c}}}}\mathrm{d}$$

$$Z_{\#,\wedge} = Z_{\#,\wedge\mathrm{a}(} + \delta\mathrm{F}a_{\#,\wedge\mathrm{a}(}\mathrm{H}$$

The $\lceil x \rceil$ operator is to round down the value x to the nearest multiple of 0.05. Logit ($Z_{i,T}$) was directly increased with an ad-hoc function $\delta(a_{i,T-1})$ that models the growth of the latent proficiency when the learning activity $a_{i,T-1}$ was taken by simulated student i at time T-1. The function δ was defined by the learning rate, the effectiveness of the instructional element taken, and (when the learning activity was an assessment quiz) the correctness of a quiz answer.

We assumed that simulated students' learning was facilitated more (i.e., a greater increase in logit) when they took effective instructional elements than ineffective elements. We also assumed that students learned more by answering a quiz correctly than incorrectly. For example, when a simulated student with a high learning rate watched an effective video, the logit was increased by 0.35, but only by 0.15 when an ineffective video was watched. For a simulated student with a low learning rate, the logit was increased by 0.31 and 0.11 respectively for effective and ineffective videos.

To control learning rate, five types of simulated students were created with different learning rates. They were labeled from R1 (the highest learning rate) to R5 (the lowest). In the simulation, 20% of simulated students were R1, 30% R2, 20% R3, 20% R4, and 10% R5—roughly reflecting a slightly skewed student population.

Under these assumptions, simulated students' learning trajectories were randomly generated. For each quality of courseware (H, M, and L), 100 instances of mock courseware were created with 1,000 simulated students. Each of the learning trajectory datasets was then converted into a learning trajectory graph (LTG). As a result, 300 LTG's were created, 100 each for H, M, and L courseware. In an LTG, Action History was encoded as a 27-bit binary vector (3 types of instructional elements, 9 each); and the Mastery Level is a decimal number (a multiple of 0.05). The latent proficiency described above was used as an estimate for Mastery Level (instead of actually applying a student model technique).

For each of the 300 LTG's, the value iteration technique was applied to compute a converse policy. As a result, 300 sets of converse policy were created, each suggesting which instructional elements were ineffective on the corresponding online courseware. Note that this simulation study models a large scale field trial with real students as if 300 instances of online courseware were tested each with 1,000 students participating. After these trials, the Rafine makes a recommendation for refinement for each instance of the courseware.

RESULTS

For the following analysis, we first evaluate the accuracy of a converse policy. We then discuss the accuracy of recommendation, which by definition is a subset of all instructional elements on the given online courseware that Rafine identifies as ineffective.

Overall Accuracy of the Converse Policy

We first evaluate the overall accuracy of a converse policy as a predictor of ineffective instructional elements. Notice that if there are *N* states in a learning trajectory graph (LTG), there are *N* converse policies generated. The accuracy of the converse policy is a ratio *n*/*N* where *n* is the number of states on which an ineffective instructional element is suggested as a converse policy.

To understand the value added by the converse policy, the chance ratio of *courseware* is defined as a ratio of ineffective to a total number of instructional elements on each courseware—e.g., for L (low) courseware, it is 16/18 = 0.89. The chance ratio of *state* is also defined among instructional elements appearing on outgoing edges of a given state as *a*/*b* where *a* is the unique number of ineffective instructional elements and *b* is the total number of unique instructional elements. In the following analysis, states where the chance ratio is equal to 1.0 or 0.0 were excluded (i.e., instructional elements on the outgoing edges were all ineffective or all effective).

Table 1: Overall accuracy of the converse policy averaged across 100 datasets for each type of courseware.

Quality	L	M	H
Chance Ratio	0.89	0.56	0.11
Accuracy	0.83	0.79	0.72

Table 1 shows the mean accuracy of a converse policy aggregated across 100 datasets for each quality of courseware. The overall accuracy of a converse policy was 0.72 even for the courseware H where only 11% (2 out of 18) of instructional elements were ineffective. These results imply that the converse policy has a high potential to accurately detect ineffective instructional element.

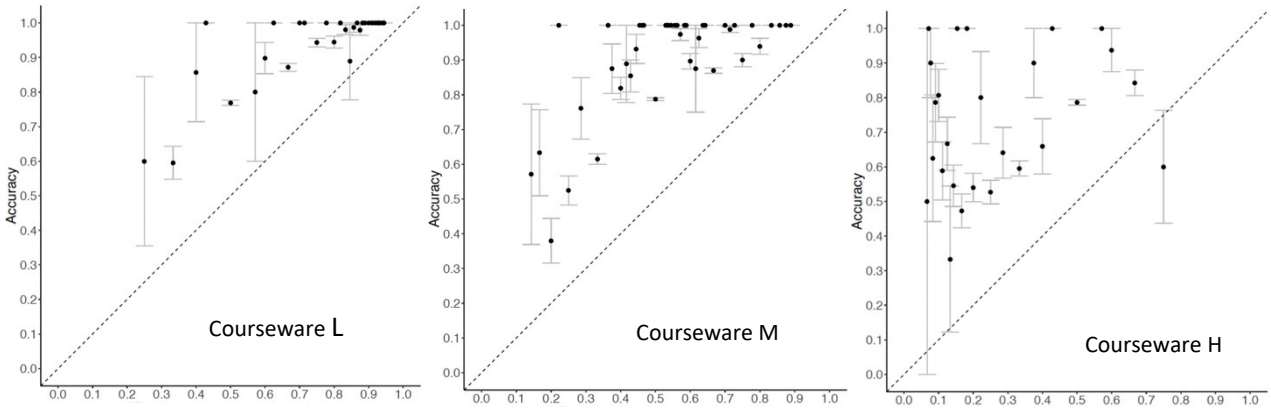


Figure 1: The accuracy of converse policy relative to states with the same chance ratio.

We hypothesize that the accuracy of a converse policy is correlated with a chance ratio of state—i.e., if a state has many outgoing edges that correspond to ineffective instructional elements, the value iteration would likely pick one of them as a converse policy. To test this hypothesis, we plotted an accuracy of a converse policy relative to a group of states with the same chance ratio as shown in Figure 1. In the figure, each data point represents a set of states that

have the same chance ratio as indicated on the x-axis. The y-axis shows the mean accuracy of a converse policy for a corresponding group of states—i.e., the ratio of states where an ineffective instructional element was selected as the converse policy to the total number of states in the group. The 45-degree line shows the chance rate. In the figure, states where the chance ratio is equal to 1.0 or 0.0 were excluded. Figure 1 indicates that the *converse policy can discriminate ineffective instructional elements from effective instructional elements far better than chance for any state in a given LTG*.

Although the converse policy can detect an ineffective instructional element at each state with a high accuracy, there are normally a notably large number of states in an LTG, so all instructional elements are included in the converse policy. Therefore, filtering the converse policy is essential for Rafine to make an actual recommendation. As our central hypothesis states, we conjecture that the frequency of being selected as a converse policy is a key for the filtering. The next section shows the accuracy of the judgement of recommendations for which instructional elements must be replaced based on the frequency heuristic.

Accuracy of Recommendations for Iterative System Improvement

We first tested if the frequency of being selected as a converse policy can be used as a filtering criterion to detect ineffective instructional elements among the converse policy. The average frequency of each instructional element being selected as a converse policy was computed by aggregating frequency values across 100 datasets. On average, each *ineffective* instructional element was selected as a converse policy 28.2 times in L, 30.6 in M, and 33.0 in H per dataset whereas each *effective* instructional element was selected 8.6 times in L, 10.0 in M, and 11.5 in H. The difference between ineffective and effective instructional elements was statistically significant for all three qualities of courseware: for L, $t(99) = 84.67$, $p < 0.05$; for M, $t(99) = 98.18$, $p < 0.05$; for H, $t(99) = 37.71$, $p < 0.05$. *These results suggest that frequency can be used as a filter to indicate ineffective instructional elements among a converse policy*.

The above observation implies that we should be able to find a frequency cut-off to determine which instructional elements must be classified as ineffective. We shall call this heuristic as the *frequency heuristic*. The question is how the cut-off should be determined, but it is rather an empirical call. We therefore compared two different cut-off thresholds—mean ± standard deviation (M±SD). The mean and the standard deviation of the frequency that individual instructional elements were selected as a converse policy were computed. Those instructional elements that appeared as a converse policy more than the cut- off are considered as ineffective. Further analysis revealed that that when the quality of courseware is low (L) to moderate (M), the M–SD cut-off yields better recall and precision than the M+SD cut-off; F1=2*precision*recall / (precision + recall) = 0.96 and 0.75 for L and M respectively with M–SD, whereas F1 = 0.38 and 0.58 with M+SD. However, when the quality of courseware is high (H), the M+SD cut-off outperforms M–SD; F1 = 0.65 for M+SD vs. 0.20 for M–SD. This implies that *at the beginning of the iterative courseware engineering, the M–SD cut-off is better, but as the courseware gets improved, the M+SD cut-off should be used*. We would want to detect as many inefficient instructional elements as possible even at a cost of false positives (i.e., the machine suggests refining even effective instructional elements).

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

We found that when students' learning trajectories were converted into a learning trajectory graph, computing the worst policy (i.e., the converse policy) using the value iteration, a well-known reinforcement learning technique, provides us with a strong clue for the effectiveness of instructional elements used in the online courseware. The converse policy is a collection of a state-action pair showing the worst action (i.e., the least effective instructional element) to be taken at a certain state. Since the number of states in a given learning trajectory is very large, all instructional elements appear in the converse policy. The frequency heuristic then differentiates those that are highly likely ineffective instructional elements from others. The proposed method, RAFINE, provides online courseware developers with an evidence- based recommendation to iteratively improve the courseware content.

The current work is a step toward realizing a fully-autonomous, self-improving online courseware— machine identifies issues and human fixes them. As for the current state of the art, we recommend GIFT to provide us API for RAFINE to give feedback to courseware developers on the quality of the individual instructional element. One idea is to flag instructional elements that are identified to be ineffective on the authoring tool GUI while the developer is editing the content. Another idea is to provide a courseware developer's dashboard that shows a birdview of courseware elements with an annotation for their predicted effective.

To yield a better prediction, RAFINE must be fed a learning trajectory graph that contains diverse learning activities. The GIFT online courseware therefore should provide students with a decent flexibility on selecting learning activities on their own.

REFERENCES

Baker, F. (2001). *The Basics of Item Response Theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Barnes, T., & Stamper, J. C. (2008). Toward automatic hint generation for logic proof tutoring using historical student data. In B. P. Woolf, E. Aïmeur, R. Nkambou & S. Lajoie (Eds.), *Proceedings of International Conference on Intelligent Tutoring Systems* (pp. 373-382): Springer.

Beck, J. E., Woolf, B. P., & Beal, C. R. (2000). ADVISOR: A machine learning architecture for intelligent tutor construction *Proceedings of the Seventeenth National Conference on Artificial Intelligence* (pp. 552-557).

Chi, M., Koedinger, K. R., Gordon, G., Jordan, P., & VanLehn, K. (2011). Instructional factors analysis: A cognitive model for multiple instructional interventions. In J. Stamper, Z. Pardos, M. Mavrikis & B. M. McLaren (Eds.), *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 61-70).

Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). An Evaluation of Pedagogical Tutorial Tactics for a Natural Language Tutoring System: A Reinforcement Learning Approach. *International Journal of Artificial Intelligence in Education, 21*(1-2), 83-113.

Clark, R., & Mayer, R. E. (2003). *e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning*. San Francisco, CA: John Wiley & Sons.

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

Fishman, B., Marx, R. W., Blumenfeld, P., Krajcik, J., & Soloway, E. (2004). Creating a Framework for Research on Systemic Technology Innovations. *The Journal of the Learning Sciences, 13*(1), 43-76. doi: 10.2307/1466932

Iglesias, A., Martinez, P., Aler, R., & Fernandez, F. (2009). Reinforcement learning of pedagogical policies in adaptive and intelligent educational systems. *Knowledge-Based Systems, 22*(4), 266-270.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science, 36*, 757-798. doi: 10.1111/j.1551-6709.2012.01245.x

Lerís, D., Sein-Echaluce, M. L., Hernández, M., & Bueno, C. (2017). Validation of indicators for implementing an adaptive platform for MOOCs. *Computers in Human Behavior, 72*, 783-795. doi: 10.1016/j.chb.2016.07.054

Martin, K. N., & Arroyo, I. (2004). Agentx: Using reinforcement learning to improve the effectiveness of intelligent tutoring systems. In J. C. Lester, R. M. Vicari & F. Paraguaçu (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 564-572).

Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2015). Faster Teaching via POMDP Planning. *Cognitive Science*, 1-43.

Shapiro, H. B., Lee, C. H., Wyman Roth, N. E., Li, K., Çetinkaya-Rundel, M., & Canelas, D. A. (2017). Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers. *Computers & Education, 110*, 35-50. doi: http://dx.doi.org/10.1016/j.compedu.2017.03.003

Shen, S., & Chi, M. (2016). Reinforcement Learning: the Sooner the Better, or the Later the Better? *Proceedings of the International Conference on User Modeling and Adaptive Personalization* (pp. 13-17).

Slavich, G., & Zimbardo, P. (2012). Transformational Teaching: Theoretical Underpinnings, Basic Principles, and Core Methods. *Educational Psychology Review, 24*(4), 569-608. doi: 10.1007/s10648-012-9199-6

Tetreault, J. R., Bohus, D., & Litman, D. J. (2007). Estimating the reliability of mdp policies: a confidence interval approach *Proc. of HLT-NAACL* (pp. 276-283).

Wiering, M., & van Otterlo, M. (Eds.). (2012). *Reinforcement Learning*. Heidelbereg, Berlin: Springer.

ABOUT THE AUTHORS

Dr. Noboru Matsuda *is an Associate Professor in the Department of Computer Science at North Carolina State University, and the director of the Innovative Educational Computing Laboratory. He leads the NSF funded project on the data-driven learning engineering methods to build adaptive online courseware where the research team investigates scalable AI techniques towards evidence-based learning engineering to build adaptive online courseware.*

Machi Shimmei *is a PhD student in the Department of Computer Science at North Carolina State University. Her research focus is on the innovation and application of cutting-edge technologies to facilitate research on education and the science of learning.*

Multimodal Machine Learning in Adaptive Instructional Systems: A Survey

Nathan Henderson, Jonathan Rowe, and James Lester
North Carolina State University

INTRODUCTION

The development and evaluation of multimodal machine learning approaches is an ongoing research area that has been the subject of growing interest in recent years. By emulating human sensory perception using multiple concurrent data channels, or "modalities," multimodal machine learning has shown promise for a range of domains related to education and training, particularly in adaptive instructional systems (AISs). Multimodal machine learning has been shown to yield improved models compared to unimodal methods, particularly in the area of affect detection and classification (Baltrušaitis, Ahuja, & Morency, 2018; Grafsgaard et al., 2014).

Recent advances in sensor technologies have enabled a growing list of applications of multimodal machine learning to different sensor-based modalities, including eye gaze data, facial expression, speech, posture, gesture, electrodermal activity (EDA), and electroencephalography (EEG). These data streams are complementary to sensor-free modalities, such as keystroke date, mouse movements, and interaction trace log data. Multimodal machine learning has been used for tasks such as automated classification and identification of affective states, including frustration, boredom, and engagement. Multimodal systems have also been devised to induce computational models for assessment (Grafsgaard et al., 2014) and metacognition (Azevedo & Aleven, 2013). Data for training multimodal machine learning models in AISs has been collected in a number of different environments, including laboratory (Taub et al., 2017), classroom (Bosch et al., 2016), and military training settings (DeFalco et al., 2018).

Multimodal machine learning shows significant promise for enabling personalized support functionalities to enhance learning outcomes and engagement in AISs. However, multimodal AISs raise challenges as well. Sensors with high sample rates generate large volumes of data to be filtered and processed, raising issues of data storage, computational resources, scalability, and modality interdependence. Sensors that rely on external hardware can also break or fail, raising issues of noise, data loss, calibration, mistracking, and interference. The inclusion of multiple parallel data streams requires each modality to be aligned and represented in a way that is compatible with a chosen multimodal machine learning algorithm (Baltrušaitis, Ahuja, & Morency, 2018). Additionally, multimodal machine learning is ideally configured to take advantage of multi-dimensional information available across the various modalities; otherwise, a multimodal approach is unlikely to be any better than an ensemble of unimodal models.

In recent years, the Generalized Intelligent Framework for Tutoring (GIFT) has emerged as an important testbed for the development and deployment of AISs. GIFT is a service-oriented framework of software tools and methods designed to streamline the process of designing, developing, and deploying AISs. Notably, GIFT provides built-in support for collecting multimodal data during student interactions with an AIS. This is enabled by the GIFT Sensor Module, which provides a configurable interface to several hardware sensors, including webcams, motion-tracking cameras, and EDA bracelets. However, much of this support is focused on collection of multimodal data. Significant gaps exist in available tools and support for the development and implementation of multimodal machine learning systems for AISs, including tools for multimodal data preprocessing, modeling, and analysis. GIFT does include some integration with existing data mining toolkits, such as RapidMiner (Mierswa, Wurst, Klinkenberg, & Scholz, 2006).

However, significant programming effort is required to utilize these features, and there is limited support for many prominent machine learning toolkits.

In this paper, we provide an overview of recent research on applications of multimodal machine learning in AISs. We describe machine learning techniques that have been used in different multimodal AISs, as well as non-instructional systems that raise parallel challenges. We discuss issues such as data fusion, data imputation, and data alignment, along with relevant algorithms. Practical considerations for multimodal data collection and analysis are noted as well. Finally, we offer several suggested directions for future enhancements to GIFT to facilitate development and utilization of multimodal machine learning in AISs, especially those developed with GIFT authoring tools.

OVERVIEW OF MULTIMODAL MACHINE LEARNING

Multimodal machine learning has its roots in audio-visual speech recognition (Baltrušaitis, Ahuja, & Morency, 2018), but with recent advances in sensor technologies and computational resources, multimodal machine learning has expanded to a wide variety of roles. In AISs, a common application of multimodal machine learning is affect detection. Affect serves a key role in shaping learning outcomes (Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2014). By devising computational models that dynamically measure learner affect at run-time, it is possible to detect and intervene in negative affective states, such as frustration and boredom, to enhance student learning outcomes (DeFalco et al., 2018; Harley, Bouchet, Hussain, Azevedo, & Calvo, 2015). Multimodal machine learning has also been utilized to predict positive affective states that correlate positively with student learning, such as engagement (Grafsgaard et al., 2014).

Sensor-Based Multimodal Affect Detection

Sensor-based multimodal systems have seen a significant increase in usage in recent years, primarily due to their inherent generalizability across a multitude of domains. This is attributable to rapid decreases in the cost and size of many sensors. Many types of sensors no longer require purchase of specialized hardware, and instead are widely available through universal platforms such as built-in webcams, microphones, eye trackers, and motion-tracking cameras like the Microsoft Kinect. Because these sensors are free of specialized hardware restrictions, they can offer a more cost-effective alternative to more expensive input channels. In a survey of multimodal affect detection systems, D'Mello & Kory (2014) observed a large number of affect detection models and detailed contemporary trends. They observed that facial expression and voice were the most commonly used modalities, occurring in over 75% of observed studies. They also noted that other sensor-based inputs such as posture, body movement, and other physiological modalities were individually present in at least 10% of studies (D'Mello & Kory, 2014).

We review a number of recent works involving sensor-based multimodal affect detection systems and observe their respective methodologies for classification, as well as their utilized modalities. Harley et al. (2015) captured multimodal data from 67 undergraduate students engaged in MetaTutor, an adaptive science-based learning environment. They captured facial expressions using a webcam in conjunction with automatic facial expression recognition software (FaceReader 5.0). They also measured physiological arousal using electrodermal activity (EDA) data using an Affectiva Q-Sensor bracelet and analyzed trends in the two modalities in conjunction with learners' self-reported affective states. Their findings indicated high level of agreement between facial expression and affective states, but low correlation between the EDA modality and affective state (Harley et al., 2015). In a similar fashion, Cooper, Arroyo, and Woolf (2011) used EDA data alongside posture and facial tracking to investigate learner engagement with Wayang Outpost, a mathematics intelligent tutoring system. They also utilized a mouse sensor that captured grip pressure. They utilized stepwise linear regression to detect a series of affective states (confident, excited, interested, or frustrated). Grafsgaard et al. (2012) utilized data from a Microsoft Kinect sensor to identify frustration, focused attention, decreased involvement, and disengagement in students interacting with a computer-mediated tutoring system for introductory Java programming. Multiple postural features were correlated with different affective states and student learning outcomes. In general, it appeared that the more a user shifted their overall posture, the less engaged and more frustrated they were.

Facial expression and posture have been investigated alongside EDA data and mouse pressure in work by Arroyo et al. (2009) aimed at classifying learner confidence, frustration, excitement, and interest among students engaged with an AIS designed for teaching geometry (Arroyo et al., 2009). Using stepwise linear regression, their work indicated that increased mouse pressure was correlated with rising frustration levels, and facial expression was indicative of approximately 60% of students' instances of affect. In addition to facial expression and posture data, Grafsgaard et al. (2014) analyzed textual dialogues between a tutoring system and student to analyze engagement, frustration, and normalized learning gains in students. Bosch et al. (2016) used clustering and Bayes Nets to construct binary classifiers for boredom, confusion, delight, engagement, and frustration. Their experiment took place in a classroom environment, where students were engaged with Physics Playground, an educational game about qualitative physics, using facial expression, head and torso positioning, and gross body movement as input modalities. Eye gaze tracking has also effectively been utilized as an indicator of learning outcomes, such as work by Rajendran, Carter, and Levin (2018) that used this modality to train a gradient tree boosting algorithm to model students' reading performance.

Additional modalities have also been investigated for classifying learners' affective states. However, the sensors required to capture these types of data are often more intrusive than facial expression or posture analysis sensors, such as webcams and motion-tracking cameras. Many multimodal systems that leverage biometric-based modalities have been devised for environments outsides of educational settings. EEG data has been used alongside Kinect data for biometric identification tasks using K-nearest neighbor clustering with histogram-oriented gradient features (Rahman & Gavrilova, 2017). EEG, EDA, and EMG modalities were modeled using Naïve Bayes classifiers, support vector machines (SVMs), and J48 decision trees for the purpose of identifying individuals' levels of arousal and valence while watching online videos (Girardi, Lanubile, & Novielli, 2017). Results show that EEG and EDA yielded the highest classification rates for arousal when used with an SVM, while all three modalities produced the highest classification rate for valence. Similarly, EDA and EEG have been simultaneously utilized to detect stress levels and cognitive load among visually-impaired people navigating an unfamiliar environment (Kalimeri & Saitis, 2016). The classifier investigated in this experiment was a random forest model. Soleymani et al. explored the use of EEG data with facial expression data for the approximation of valence and arousal levels of students watching a series of emotion-invoking video clips. They used long short-term memory recurrent neural networks (LSTM-RNN) and continuous conditional random fields for their classification models. Their results indicated that facial expression data was inherently more informative than EEG data for their task, and a majority of the EEG features were a result of facial expression contamination. However, the EEG modality was beneficial when used in a complimentary role alongside the facial expression modality (Soleymani, Asghari-Esfeden, Fu, & Pantic, 2016).

Multimodal Deep Learning

Deep learning techniques, such as LSTM-RNNs, have seen a huge increase in interest in recent years, particularly due to significant improvements in computational hardware such as graphical processing units. In a survey paper focused on deep multimodal learning, Ramachandram and Taylor (2017) attribute increased interest in deep learning to its ability to form a hierarchical representation of each modality simultaneously, offering a distinct advantage over unimodal classifiers. Neural network architectures such as convolutional neural networks, autoencoders, LSTM-RNNs, and feedforward neural networks have also been shown to serve as effective multimodal machine learning models for tasks such as affect detection, sentiment analysis, image annotation, and speech classification (Ramachandram & Taylor, 2017). Common modalities for these solutions include audio-visual information, text, speech/dialogue data, and optical flow. Deep feedforward neural networks have been shown to yield improved performance in tasks such as frustration detection over non-neural models, such as SVMs (Henderson et al., 2019), which may be attributable to their innate ability to learn complex relationships across high-dimensional data as well as their capacity to process data while keeping spatial and temporal context intact (Pei, Yang, Jiang, & Sahli, 2015).

CHALLENGES IN MULTIMODAL MACHINE LEARNING

Although multimodal machine learning shows significant promise within AISs, there are still a multitude of risks and issues to be addressed, particularly when dealing with sensor-based models. Some concerns are raised when

dealing with the dimensionality of the data itself. For example, multimodal systems often require high-dimensionality data, which can improve performance but also increase computational workloads and hardware constraints. Another issue is the spatial nature of many modalities. Although temporal information in different modalities has been demonstrated to be beneficial to multimodal classifiers (Henderson et al., 2019), many sensors only capture positional or spatial data at discrete time points; temporal context is not explicitly recorded. Other hardware-related issues can also arise, yielding significant noise or the loss of an entire data channel altogether. Because multimodal machine learning often involves the creation of a singular model for a multitude of data streams, it becomes imperative that each modality is equally accessible and interpretable. This raises issues of data alignment and representation, as well as calibration of the sensors themselves. In this section, we discuss several common issues and obstacles that are raised by multimodal machine learning-based systems, as well as different efforts to remedy several of these issues.

Temporal Context

As stated in the previous paragraph, sensors often capture limited temporal context about subjects. While this does not directly inhibit the creation or deployment of multimodal systems, such information can provide insight into the state of the subject captured by the sensor. For example, this was recently shown to be beneficial to the creation of a multimodal machine learning-based model for run-time affect detection in a game-based learning environment for emergency medical training (Henderson et al., 2019). A single modality containing spatial posture data (i.e., torso position) was captured using a Microsoft Kinect sensor. A second, synthetic modality that captured temporal data (i.e., torso velocity) was generated by taking the derivative of each captured instance of participants' postural positions. This temporal information improved the performance of the affect detector over a previously published baseline that utilized only the spatial Kinect features. Another example of this approach took (1) body lean angle, (2) slouch factor, (3) quantity of motion, and (4) contraction index from a single postural modality (Sanghvi et al., 2011). These input vectors were utilized for the classification of elementary school students' engagement with an automated companion in a game-based learning environment (iCat). These features served as artificial temporal modalities, an attempt to solve the issue of missing temporal context from sensor-based data. An alternative approach is to employ machine learning to derive temporal context from a multimodal dataset using techniques such as recurrent neural networks (Chen & Jin, 2015). Continuous generation models have also been used as a generative approach to preserve temporal information across modalities (Baltrušaitis, Ahuja, & Morency, 2018).

Data Preprocessing

There are several issues that arise during data preparation prior to the application of multimodal machine learning techniques. For example, a common issue in affect detection is the problem of imbalanced class labels. In many educational settings, a student is more likely to exhibit displays of concentration than frustration or surprise, which may adversely impact a classifier's ability to accurately detect certain affective states. This calls for application of oversampling techniques to training data. One method to address this problem is minority cloning. This oversamples positive instances of a sparse affective state, bringing the data to a balance that is closer to an even ratio of positive and negative instances. A more sophisticated approach that is commonly used in affect detection is Synthetic Minority Over-sampling Technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). SMOTE generates synthetic samples from the minority class based on existing samples.

An important step in many practical applications of multimodal machine learning is feature selection and feature reduction. Because each modality can have high dimensionality, it is important to eliminate redundant or irrelevant features to save computational time and resources. There are range of feature selection algorithms that have been utilized to distill feature representations in AISs. One example is forward selection: this is a greedy selection algorithm that trains a model on each feature and selects the feature whose model returns the highest performance. This process continues until a preset number of features is reached. Other feature selection methods include univariate selection, tree-based feature selection, and removing features with low variance. A common approach to feature reduction is principal component analysis (PCA). PCA involves reducing a multivariate dataset to lower-

dimensionality linearly correlated values called "principal components" (Chandrashekar & Sahin, 2014). Autoencoders have also seen use as a feature reduction method, and they are particularly useful due to their ability to decode the reduced data to its original representation (Jaques, Taylor, Sano, & Picard, 2017).

Data Imputation

Multimodal AISs, particularly sensor-based systems, face inherent risks associated with hardware failure. Physical hardware can be unreliable or inconsistent, leading to issues such as data noise, data loss, and outliers. A common consequence is missing data for one or more modalities. Another common issue is data noise within a modality. This can occur due to background activity captured by a sensor (i.e., someone walking in the background), or inconsistent behavior from the sensor itself. While a common solution is to discard data samples that contain missing or invalid data, an alternative approach is to impute missing values using the available samples. One simple approach to data imputation is *mean imputation*, which involves replacing each missing value with the mean of existing values for that particular feature. A more sophisticated method is the use of autoencoders to impute data (Jaques, Taylor, Sano, & Picard, 2017). This involves training an autoencoder with a subset of data that does not contain any missing values. This trained model is then used to approximate the missing values. This method can be effectively applied to sparse missing data, as well as entire missing modalities. Autoencoders have also been commonly used as a denoising technique for noisy or inconsistent data.

Data Fusion

Data fusion deals with the integration of data from multiple modalities for the purpose of classification or regression. Thus, it is a critical step in the creation of multimodal machine learning models (Baltrušaitis, Ahuja, & Morency, 2018). The majority of data fusion techniques are model-agnostic; that is, the data fusion is not reliant on a particular machine learning algorithm, and it occurs prior to, or after, classification or regression has taken place (Baltrušaitis, Ahuja, & Morency, 2018). These approaches are commonly divided into three categories: early fusion (feature-level), late fusion (decision-level fusion), and hybrid fusion.

Early fusion involves the concatenation of feature vectors from multiple modalities, and it occurs immediately after feature extraction. This is arguably the simplest data fusion method, since concatenation is a relatively straightforward operation and the method requires only a single machine learning model. Late fusion involves training a unimodal classifier for each modality and then fusing the resulting predictions from each classifier. Fusion can be accomplished in several possible ways, including averaging, voting, weighting, or applying another machine learning-based model (Baltrušaitis, Ahuja, & Morency, 2018). This approach does allow for different machine learning models to be used on each modality, which may increase overall model performance. This method is also more robust, allowing for models to be trained even in the presence of missing modalities. Hybrid fusion combines predictions from early fusion with additional unimodal predictors. Recent efforts to evaluate data fusion methods include work by Rahman and Gavrilova (2017), which used a form of late fusion on EEG and Kinect posture data for biometric identification. Similarly, Kalimeri and Saitis (2016) used early fusion on EEG and EDA modalities for the detection of stress levels, and Patwardhan and Knapp (2016) used a variety of body tracking modalities in combination with late fusion for the purpose of affect detection. Finally, Henderson et al. evaluated both early and late fusion techniques with a combination of spatial and temporal posture modalities for frustration detection in a game-based learning environment (Henderson et al., 2019).

There are also data fusion techniques that implicitly handle multimodal data; we refer to these as *model- based approaches* (Baltrušaitis, Ahuja, & Morency, 2018). One example of a model-based approach is multiple kernel learning models. These are an extension of SVMs, but apply the kernel-based learning approach to multiple modalities. Another alternative is probabilistic graphical models. Originally devised using hidden Markov models and Bayesian networks, probabilistic graphical models have expanded to include conditional random fields. Graphical models are useful due to their ability to process spatial and temporal features from multimodal data, and they often lead to interpretable models (Baltrušaitis, Ahuja, & Morency, 2018). As stated before, deep learning has

become a prominent method for multimodal machine learning. This is partly due to its innate ability to fuse encoded features from multiple modalities. Deep neural networks offer the ability to learn complex relationships across high-dimensionality datasets, as well as the ability to process spatial and temporal information through the use of CNNs and RNNs, respectively.

Data Alignment

Multimodal data alignment is a process that accounts for the relationships between sub-components of two or more modalities (Baltrušaitis, Ahuja, & Morency, 2018). Data alignment is particularly relevant when multiple modalities operate or sample at differing frequencies. This task can be undertaken using either explicit or implicit alignment. Explicit alignment is an alignment process that is the primary objective of a modeling analysis. Implicit alignment is an intermediate step within a larger, overarching task. Implicit alignment often involves a latent representation of the separate modalities, and it often occurs during model training. As a result of this approach, the models do not explicitly align the data, but latently align the data during the training phase.

Explicit alignment aims to increase the correlation between two modalities' components. This allocates increased emphasis on a similarity metric that evaluates the quality of the alignment between two or more modalities. One common approach is called dynamic time warping (DTW). DTW is a dynamic programming approach that can be applied to time-series data—temporal alignment is often a primary issue in multimodal datasets—particularly when multiple sensors or input channels are involved. DTW computes the similarity between two modalities and inserts additional frames within the modalities to find an optimal match (Baltrušaitis, Ahuja, & Morency, 2018). This requires timesteps between the two modalities to be comparable and compatible with the given similarity metric. More recently, canonical correlation analysis has been used as a linear transformation serving as the similarity metric for DTW. This method allows DTW to discover linear relationships across multiple modalities in the temporal dimension, but it does not work well with non-linear relationships. Deep learning techniques have also been used for data alignment, particularly to measure similarity between modalities. However, because deep neural networks are typically utilized in supervised fashion, there is an implicit requirement for pre-aligned data to be used to train deep learning models. Often, datasets lack a subset of explicitly annotated data, restricting the utility of supervised data alignment techniques (Baltrušaitis, Ahuja, & Morency, 2018).

Implicit alignment is used for tasks where explicit alignment is either not useful or feasible, such as speech recognition or machine translation. Early work in implicit alignment involved the use of graphical models. However, the usage of this method has waned over time due to the need for manual construction of the graphical mapping between modalities and the need for previously-aligned training data. More recently, deep neural networks have become a primary method of implicit alignment. Often, this takes place through the use of autoencoder models as well as cross-modality retrieval models.

MULTIMODAL MACHINE LEARNING IN GIFT

GIFT has been used to develop and deploy AISs in a range of research studies (Aleven et al., 2018; DeFalco et al., 2018; Goldberg & Cannon-Bowers, 2015). Several studies have utilized GIFT's Sensor Module to collect multimodal data. However, GIFT provides limited support for downstream analysis and modeling of multi-channel data using multimodal machine learning techniques. We offer several recommendations for potential enhancements to GIFT which would facilitate the development and deployment of AISs that leverage multimodal data streams. Currently, GIFT supports sensors for posture, gesture, facial expression analysis, EDA, and EEG. Additional data channels can be added to GIFT-based AISs by integrating additional hardware sensor types, such as electromyography and eye tracking. Further, sensor integration need not be restricted to a single sensor for each modality; it is conceivable that there would be benefits to supporting multiple sensors concurrently that focus on a single modality (i.e. multiple Microsoft Kinect sensors positioned at different locations around a learner).

Additional data preprocessing techniques could also be integrated into the GIFT Sensor Module to improve the preparation of data prior to it being sent to the Learner Module. One example is feature scaling, typically performed through data normalization or standardization, which is a step that is often necessary in machine learning analysis, particularly in deep learning. Providing solutions to address class label imbalances in recorded data could also prove useful, including support for algorithmic oversampling techniques such as SMOTE. Notably, this would only be applicable to recorded data that is labeled prior to oversampling. Another prospective enhancement is integrating feature selection and feature reduction techniques. This would enable GIFT to present the Learner Module with data that omits redundant or uninformative information that can be captured in raw sensor data. Examples include forward feature selection or elimination of features that fall below a pre-set variance threshold. Alternately, dimensionality reduction techniques, such as PCA, can be integrated as well. These techniques will decrease the amount of data processed by the Learner Module, thus reducing run-time computational requirements.

Another prospective improvement to GIFT would be the integration of data imputation methods. This would reduce the negative impact of missing data in GIFT's analysis pipeline, and it also ensures the preservation of each data point in the raw sensor data. Simple imputation methods, such as mean imputation, ensure that missing or invalid data do not adversely impact the processing pipeline, and they do not require previously-labeled data or model training. More complex methods, such as the autoencoder-based methods, impute missing data more accurately, but they require pre-existing trained models. This renders the approach ineffective in instances where a modality is missing a majority of its data. However, the trained autoencoder can also be utilized to denoise the data, which can boost classifier performance.

Data fusion techniques could also be implemented in GIFT to aid the Learner Module in dealing with multiple data channels. Feature-based data fusion (Early Fusion) is programmatically simple to integrate as it requires modalities to be concatenated prior to being passed to other modules. Alternative feature-level fusion methods have been shown to offer improvement over simple feature concatenation. For example, performing feature selection on individual modalities prior to feature concatenation has been demonstrated to improve affect classification results (Henderson et al., 2019). Decision-level fusion is more complex to implement due to the need for a decision selection schematic, as well as the need for a machine learning model for each modality.

Expanded support of machine learning models and tools would also introduce to GIFT the capability to implement an entire multimodal data processing pipeline, including initial data capture, preprocessing, imputation, and modeling. Recent years have seen growing interest in deep learning-based models in AISs for a variety of learner modeling tasks, including run-time assessment and affect detection. Enhanced support for deep neural networks, including LSTM-RNNs, as well as other ML algorithms within GIFT would provide an expanded range of modeling options. It should be noted that the addition of these ML techniques stipulates a requirement for labeled training data, as well as computational resources for data- intensive deep learning algorithms, such as RNNs.

The most significant challenge to the integration of multimodal machine learning in AISs is handling disparate data streams. This is a common problem for multimodal systems deploying sensors operating at different sampling rates and within different time intervals. For modeling techniques such as data fusion to be possible, this issue must be addressed. We recommend two types of data alignment techniques to address this problem: explicit and implicit alignment. Temporal misalignment can be explicitly handled through DTW, although this method requires that the time axis between modalities be compatible with the similarity metric in the DTW algorithm. Additionally, non-linear relationships between modalities increases the difficulty of modality alignment. Although deep learning techniques have emerged as an effective approach to implicit alignment, it requires pre-labeled training data, which is usually not readily available in multimodal AISs involving disparate data streams. Data alignment continues to be widely researched and is an area of significant interest in the development of multimodal machine learning systems.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

Multimodal machine learning will have a critical role in the design, development, and evaluation of AISs. Multimodal data streams enable the creation of data-rich models of student learning and engagement, which can be utilized to inform adaptive interventions to improve student outcomes. We have provided a survey of recent work on multimodal machine learning in AISs. We detail common machine learning techniques used to implement multimodal AISs, including key components of the multimodal data processing pipeline. Further, we detail specific challenges faced by developers of multimodal AISs, including common issues in data collection, alignment, and modeling.

GIFT has significant promise for facilitating future development of multimodal AISs. We recommend that future research and development efforts focus on integrating an expanded range of machine learning algorithms, addressing common issues raised by sensor-based AISs, and implementing solutions to data misalignment issues, particularly along the temporal dimension. By extending GIFT to include enriched multimodal machine learning capabilities, significant strides can be made to increase access to computational solutions for enhancing learner models and enabling adaptive pedagogical functionalities that improve learning outcomes and instructional effectiveness.

REFERENCES

Aleven, V., Sewall, J., Andres, J. M., Sottilare, R., Long, R., & Baker, R. (2018). Towards adapting to learners at scale: integrating MOOC and intelligent tutoring frameworks. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (Article no. 14). New York, NY: ACM.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). Emotion sensors go to school. In *Proceedings of the 14th International Conference on Artificial Intelligence In Education* (pp. 17– 24). http://doi.org/10.3233/978-1-60750-028-5-17

Azevedo, R., & Aleven, V. (2013). Metacognition and learning technologies: an overview of current interdisciplinary research. In *International handbook of metacognition and learning technologies.* (pp. 1–16). Springer, New York, NY. http://doi.org/10.1007/978-1-4419-5546-3_47

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. http://doi.org/10.1109/TPAMI.2018.2798607

Bosch, N., Mello, S. K. D., Dame, N., Dame, N., Baker, R. S., Shute, V., … Zhao, W. (2016). Detecting student emotions in computer-enabled classrooms. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (pp. 4125–4129).

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over- sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357. http://doi.org/10.1613/jair.953

Chen, S., & Jin, Q. (2015). Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* (pp. 49–56). ACM. http://doi.org/10.1145/2808196.2811638

Cooper, D. G., Arroyo, I., & Woolf, B. P. (2011). Actionable affective processing for automatic tutor interventions. In *New perspectives on affect and learning technologies* (pp. 127–140). New York, NY: Springer. http://doi.org/10.1007/978-1-4419-9625-1

D'Mello, S. K., & Kory, J. (2014). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, *47*(3), 43. http://doi.org/10.1145/2682899

DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., … Lester, J. C. (2018). Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*, *28*(2), 152–193. http://doi.org/10.1007/s40593-017-0152-1

Girardi, D., Lanubile, F., & Novielli, N. (2017). Emotion detection using noninvasive low cost sensors. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 125-130). IEEE.

Goldberg, B., & Cannon-Bowers, J. (2015). Feedback source modality effects on training outcomes in a serious game: Pedagogical agents make a difference. *Computers in Human Behavior*, *52*, 1-11.

Grafsgaard, J., Boyer, K., Wiebe, E., & Lester, J. (2012). Analyzing posture and affect in task-oriented tutoring. In *FLAIRS Conference* (pp. 438–443). Retrieved from http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS12/paper/download/4447/4843

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014). Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue. In *Proceedings of the Seventh International Conference on Educational Data Mining* (pp. 122–129). London, UK: International Educational Data Mining Society. http://doi.org/10.1182/blood-2013-10-529982.The

Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014). The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the Sixteenth ACM International Conference on Multimodal Interaction* (pp. 42–49). ACM. http://doi.org/10.1145/2663204.2663264

Harley, J. M., Bouchet, F., Hussain, M. S., Azevedo, R., & Calvo, R. (2015). A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior*, *48*(May), 615–625. http://doi.org/10.1016/j.chb.2015.02.013

Henderson, N. L., Rowe, J. P., Mott, B. W., Brawner, K., Baker, R. S., & Lester, J. C. (2019). 4D Affect Detection : Improving Frustration Detection in Game-Based Learning with Posture-Based Temporal Data Fusion. In *Proceedings of The 20th International Conference on Artificial Intelligence in Education (in press)*.

Jaques, N., Taylor, S., Sano, A., & Picard, R. (2017). Multimodal Autoencoder : A Deep Learning Approach to Filling In Missing Sensor Data and Enabling Better Mood Prediction, 0–6.

Kalimeri, K., & Saitis, C. (2016). Exploring multimodal biosignal features for stress detection during indoor mobility. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 53–60). ACM. http://doi.org/10.1145/2993148.2993159

Mierswa, I., Wurst, M., Klinkenberg, R., & Scholz, M. (2006). Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 935–940). http://doi.org/10.1080/14672715.1968.10405148

Patwardhan, A., & Knapp, G. (2016). Multimodal affect recognition using Kinect. *ArXiv Preprint ArXiv:1607.02652*. Retrieved from http://arxiv.org/abs/1607.02652

Pei, E., Yang, L., Jiang, D., & Sahli, H. (2015). Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 208–214). IEEE. http://doi.org/10.1109/ACII.2015.7344573

Rahman, W., & Gavrilova, M. L. (2017). Emerging EEG and Kinect face fusion for biometric identification. In *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). IEEE. Rajendran, R., Carter, K. E., & Levin, D. T. (2018). Predicting Learning by Analyzing Eye-Gaze Data of Reading Behavior. *International Educational Data Mining Society*.

Ramachandram, D., & Taylor, G. W. (2017). Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108. http://doi.org/10.1109/MSP.2017.2738401

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th International Conference on Human-robot Interaction* (pp. 305–312). ACM. http://doi.org/10.1145/1957656.1957781

Soleymani, M., Asghari-Esfeden, S., Fu, Y., & Pantic, M. (2016). Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *IEEE Transactions on Affective Computing*, *7*(1), 17–28. http://doi.org/10.1109/TAFFC.2015.2436926

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with CRYSTAL ISLAND. *Computers in Human Behavior*, *76*, 641–655. http://doi.org/10.1016/j.chb.2017.01.038

ABOUT THE AUTHORS

Nathan Henderson *is a PhD student in Computer Science at North Carolina State University. He received his BS degree in Electrical and Computer Engineering from Auburn University and his MS degree in Computer Science from The University of Alabama in Huntsville. His research focuses on multimodal machine learning techniques for advanced learning technologies. Prior to graduate school, he was employed as a software developer for multiple U.S. Department of Defense contractors, where he supported several machine learning-related contracts.*

Dr. Jonathan Rowe *is a Research Scientist in the Center for Educational Informatics at North Carolina State University, as well as an Adjunct Assistant Professor in the Department of Computer Science. He received the PhD and MS degrees in Computer Science from North Carolina State University, and the BS degree in Computer Science from Lafayette College. His research focuses on the intersection of artificial intelligence and human-computer interaction for advanced learning technologies, with an emphasis on game-based learning environments, intelligent tutoring systems, multimodal analytics, learner modeling, and computational models of interactive narrative generation.*

Dr. James Lester *is Distinguished University Professor of Computer Science at North Carolina State University, where he is Director of the Center for Educational Informatics. His research centers on transforming education with technology-rich learning environments. With a focus on adaptive learning technologies, his research spans intelligent tutoring systems, game-based learning environments, affective computing, and tutorial dialogue. The adaptive learning environments he and his colleagues develop have been used by thousands of students in K-12 classrooms. He received his PhD in Computer Science from the University of Texas at Austin in 1994. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI).*

Understanding Novelty in Reinforcement Learning-Based Automated Scenario Generation

Jonathan Rowe, Andy Smith, Randall Spain, and James Lester
North Carolina State University

INTRODUCTION

Simulations will serve a critical role in the next generation of training. A key feature of simulation-based training is the capacity to deliver scenarios that support the acquisition, practice, and assessment of domain-critical knowledge, skills, and abilities. The recently formed Close Combat Task Force, ordered by the U.S. Secretary of Defense, has called for Soldiers to take part in 25 simulated battles before facing their first real contact (Judson, 2018). To attain this level of training, devising effective methods for the creation and delivery of simulation-based training scenarios is essential. However, creating training scenarios for simulation-based environments poses significant challenges: authoring simulation-based training scenarios is often resource-intensive; it requires specialized knowledge about specific simulators and authoring tools; scenarios often support only limited reuse; and most scenarios adhere to a one-size-fits-all approach that does not support adaptivity, either with respect to the changing needs of instructors or changing needs of trainees. These issues point toward the need for automated scenario generation to increase the availability and diversity of training scenarios across a range of tasks, domains, and simulation environments.

A key criterion in evaluating the effectiveness of automated scenario generation systems is their capacity to create training scenarios that are novel. Novel training scenarios are (a) meaningfully different from previously experienced scenarios, and (b) aligned with relevant instructional objectives for training. For example, it is possible to generate a vast range of variations on a training scenario by subtly adjusting the location of a single entity in a simulated environment. However, these variations could hardly be considered novel. Similarly, generating training scenarios that are misaligned with relevant instructional objectives, or unrealistic with respect to real-world scenarios, is of little value as well. Instead, novel scenarios should differ from existing scenarios in ways that are pedagogically meaningful—for example, they modify a scenario's difficulty or alter the format of a correct solution—in order to provide learners with new, beneficial training opportunities.

In this paper, we explore the role of novelty in data-driven automated scenario generation for simulation-based training environments. This is informed by our ongoing work to develop DEEPGEN, a reinforcement learning (RL) framework for automated scenario generation in the domain of Call for Fire (CFF) training with Virtual Battlespace 3. DEEPGEN utilizes RL techniques to induce computational models for run-time tailoring of scenarios to achieve instructor-specified training objectives (Rowe, Smith, Pokorny, Mott, & Lester, 2018). Specifically, we formalize scenario generation as an RL task that involves sequential decision making about enacting adaptations (i.e., actions) to an exemplar scenario, observing learner interactions with the generated scenario (i.e., trajectories), and using the resultant learner performance data (i.e., reward) to refine an internal decision-making model for future scenario adaptation decisions.

To guide our discussion of novelty, we draw upon the psychology literature on creativity. Kaufman and Beghetto (2009) devised the Four C Model of Creativity, which distinguishes between *Big-C*, *Pro-c*, *Little-c* and *mini-c* conceptualizations of creativity. The Big-C category refers to eminent creativity, which is understood as creative work that is historically significant, lasting, and signifying creative genius. Pro-c creativity refers to effortful progression toward, and often antecedent to, Big-C status; it is associated with professional-level expertise that exceeds novice-level creativity, but does not yet reach a level of Big-C contribution. Little-c creativity is everyday focused, referring to expressions of creativity that are performed by non-experts, such as inventive problem solving or creative endeavors undertaken as a hobby. Mini-c creativity is a signature of the learning process; it is defined as the "novel and personally meaningful interpretation of experiences, actions, and events" (Kaufman & Beghetto, 2009).

Drawing upon the Four C Model, we conceptualize novelty in automated scenario generation in terms of four categories. First, we distinguish training scenarios that are wholly new and unique, and valuable; this is akin to anticipating future scenarios that have never been encountered before. Second, we distinguish scenarios that are new to an instructor (or simulation environment), but are not necessarily novel in a universal sense. Generating scenarios in this category is a significant enhancement to the training capacity of a simulation-based training environment. Third, we distinguish training scenarios that are new to a particular learner, or group of learners, even if they were already pre-existing. Fourth, we distinguish novel experiences that a learner might have with a scenario that they have previously experienced; successfully completing a scenario for the first time would fall into this category.

Utilizing this conceptual model, we analyze a prototype version of the DEEPGEN scenario generation system in terms of the Scenario Adaptation Library that was developed for CFF training, as well as the scenarios that can be automatically generated by the prototype system. We discuss novelty in terms of several dimensions of CFF scenario adaptation, including adjustments to low-level features such as unit types, terrain, weather, and entity locations, as well as higher-level features, such as adversary behaviors, mission objectives, and training contexts. We explore how novelty can be operationalized within an RL framework for automated scenario generation. Finally, we discuss the implications for the Generalized Intelligent Framework for Tutoring (GIFT) as they relate to the integration of automated scenario generation capabilities within adaptive training systems. Developing theoretically grounded approaches to conceptualizing novelty in automated training scenario generation will be critical to meeting the mandate of next-generation simulation-based training.

RELATED WORK

We approach automated generation of training scenarios from the perspective of a related research area: interactive narrative generation (Riedl & Bulitko, 2012). Interactive narrative generation focuses on the design of computational models for dynamically generating and tailoring digital interactive experiences in which users drive an unfolding storyline through their own actions and decisions. A range of computational techniques have been investigated for interactive narrative generation, including classical AI planning (Young et al., 2013), adversarial search (Nelson & Mateas, 2005), case-based reasoning (Fairclough, 2004), and machine learning (Wang et al., 2018). Grounded in this work, we conceptualize scenarios in terms that are analogous to interactive narrative systems: scenarios consist of sequences of events that unfold within a virtual environment. A scenario specification includes the initial state of the virtual world, including its terrain, agents, buildings, weather, and overall task instructions presented to the user. In addition, the set of agent behaviors and associated triggers that define how events play out within the virtual environment are integral to the scenario. In a training context, scenarios specify a set of learning objectives to be addressed as well as assessment criteria. Finally, scenarios are completable, and they should be both coherent and internally consistent.

Recent years have seen growing interest in the use of machine learning techniques for data-driven automated scenario generation in education and training. This includes applications of dynamic decision networks (Mott & Lester, 2006), dynamic Bayesian networks (Lee, Rowe, Mott, & Lester, 2014), and reinforcement learning techniques (Rowe & Lester, 2015; Wang et al., 2018). However, much of this work has focused on devising computational models for tailoring narrative-centered learning experiences to ensure they are effective and engaging; there has been comparatively little work investigating novelty in machine learning-based frameworks for automated scenario generation. Although novelty has been investigated for scenario generation with evolutionary algorithms (Folsom-Kovarik & Brawner, 2018), to date, novelty has not been a central factor in machine learning-based approaches.

DEEPGEN RL-BASED SCENARIO GENERATION FRAMEWORK

To contextualize our discussion of novelty in automated scenario generation, we cite examples from an ongoing project in our lab that investigates the design and development of a data-driven scenario generation system,

DEEPGEN. DEEPGEN formulates automated scenario generation as an instance of data-driven interactive narrative generation utilizing deep reinforcement learning techniques.

To serve as an initial testbed, DEEPGEN focuses on automated scenario generation in the task domain of artillery call-for-fire (CFF) training. Broadly speaking, a CFF mission consists of an infantry soldier requesting indirect fire on a target from supporting artillery (e.g. field artillery, unmanned aircraft). The requesting soldier, or forward observer (FO), follows a defined communication protocol to identify himself, describe the mission type, describe the target and location, and describe the method of engagement. The mission continues as the FO may choose to adjust fire as necessary based on the results of initial shots, and conclude the mission by relaying a battle damage assessment once the target has been hit. Given this general structure, there are a broad range of scenario adaptations that can be enacted to augment the difficulty of a CFF training scenario, such as changing the type of mission, modifying enemy behaviors, modifying weather and time-of-day, changing the types and locations of targets, and varying the types of equipment in the FO's loadout.

Table 1. Partial Scenario Adaption Library for CFF domain

Adaptable Elements of Scenario	Scenario Adaptation Variants
Target Type	• Infantry squad • Transport vehicle • Tank (T72)
Target Behavior	• Stationary • Patrol • Move to waypoint
Target Reaction To Fire	• No reaction • Stop movement • Flee to cover • Return to base

As dynamic scenario adaptation involves enacting a series of decisions about how to orchestrate training events at run-time, DEEPGEN enumerates the full range of possible adaptations in a *Scenario Adaptation Library.* The Scenario Adaptation Library defines the space of possible transformations that can be applied to example (i.e., parent) scenarios in order to produce new (i.e., child) scenarios. Thus, DEEPGEN approaches novelty through the systematic combination and application of individual scenario adaptations. In the CFF training testbed, we have defined 16 possible dimensions for scenario adaptation corresponding to more than 1,000,000 possible scenario variations that could be generated from a single example scenario. A sample of adaptations from DEEPGEN's Scenario Adaptation Library can be found in Table 1.

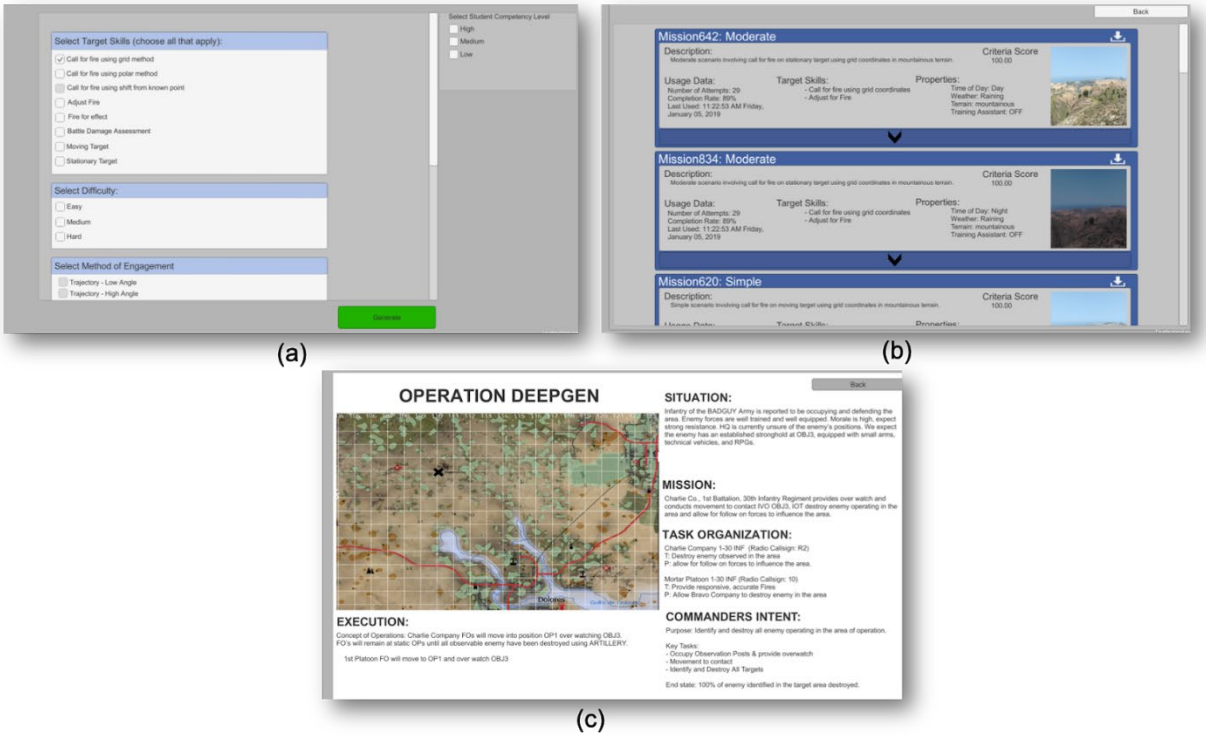


Figure 1. Screenshots of prototype DEEPGEN instructor tool.

To generate scenarios using DEEPGEN, an instructor first selects a set of learning objectives and selection criteria to guide the scenario generation process, as shown in Figure 1(a). Using these inputs, DEEPGEN systematically generates a set of scenario adaptations that can be enacted with a given parent scenario, ranking these combined variations according to their match to the selected criteria, as shown Figure 1(b). Finally, the instructor can view a more detailed view of each scenario in the form of a Warning Order, as shown in Figure 1(c), before selecting one or more scenarios to be downloaded and realized within a 3D simulation-based training environment. To serve as a testbed simulation environment for CFF training, DEEPGEN interoperates with Virtual Battlespace 3 (VBS3). Developed by Bohemia Interactive Simulations, VBS3 is a 3D simulation platform that is widely used by the U.S. Army for a range of training purposes, including IED training, land navigation, route clearance, and convoy training. Furthermore, we utilize the VBS3Fires plug-in, a third-party tool created by SimCentric Technologies that provides a GUI interface and ballistics simulation engine for training CFF in VBS3. Scenario specifications generated by DEEPGEN are realised in VBS3 by modifying example VBS missions through a semi-automated compilation process that produces a full set of executable and configuration files required by VBS.

In the current prototype version of DEEPGEN, the Scenario Adaptation Library for CFF training has been hand-authored through close collaboration with U.S. Army subject-matter experts, thus guaranteeing that generated scenarios are completable and coherent. However, there are significant overlaps between many of the generated scenarios. This raises important questions related to scenario novelty: How should we understand the degree of novelty that is supported by DEEPGEN's automated scenario generation framework? To what extent can we enhance the degree of novelty exhibited in scenarios created by DEEPGEN? And how can novelty be conceptualized to advance training objectives in task domains like CFF training?

NOVELTY IN DEEPGEN

To frame our discussion of novelty in scenario generation, we use the Four C Model of Creativity devised by Kaufman and Beghetto (2009). The Four C model extends traditional creativity research, which has historically focused mainly on "genius"-level creativity (Big-C) and "everyday" creativity (Little-c) to include two new levels,

Pro-c and Mini-c. In this section we expand on each of the Four C's, discussing how they relate to novelty in scenario generation and providing specific examples of how they are, and are not, addressed by DEEPGEN.

Big-C creativity refers to "clear-cut, eminent creative contributions" (Kaufman & Beghetto, 2009). Depending on the domain, this can refer to creative works ranging from award-winning musical compositions to scientific discoveries to Pulitzer Prize-winning novels. This category exists to distinguish exceptional creative people and works from highly competent, even professional level creators. From a scenario generation perspective, we identify Big-C level novelty as referring to scenarios that introduce fundamental, long-lasting changes to the rules or performance expectations associated with a target domain. For example, consider the case of AlphaGo, an artificial intelligence-based game-playing agent designed by Google DeepMind to automatically learn how to play the board game Go (Silver et al., 2017). AlphaGo famously performed at a level capable of beating top-rated human players, and on the 37th move of Game Two in its historic match against 18-time world champion Lee Sedol, AlphaGo performed a move that was not only effective, but that commentators later regarded as "creative," "beautiful," and likely to be incorporated into future matches by human players (Metz, 2016). It is important to note that the system did not just beat expert-level human players, but it did so by exhibiting a novel strategy that had not been previously encountered. Another example of Big-C novelty in scenario generation comes from the Millennium Challenge 2002 wargaming exercise (Borger, 2002). During the exercise, the commander of the "Red" force adopted a variety of novel strategies that effectively defeated the "Blue" forces despite their superior technological advantages. The strategies were so effective that the simulation was reset, and the rules of engagement were rewritten because the "Blue" force had been defeated so quickly. Although created by a human author, the Millennium Challenge example demonstrates a Big-C level contribution from a scenario due to its impact on the strategies employed in future war-gaming exercises.

For a system to produce Big-C level novelty in scenario generation, a variety of factors are required. By definition, these scenarios make a contribution that did not previously exist, and thus the scenario generator must have access to a broad range of flexibility and freedom to explore the space of possible scenarios. This calls for direct access to robust, high-fidelity simulation environments, or game-like environments where agents can be trained through "self-play" or other strategies likely to produce emergent behavior, which contrasts with approaches designed to mimic expert human performance. Notably, it can be difficult to recognize innovative scenarios if they are not highly effective in comparison to competing options (i.e., innovative scenarios might not be selected by an imperfect optimization process). Thus, even given ideal conditions, it is not guaranteed that any system will generate scenarios at the Big-C level of novelty.

The next level of novelty we consider is Pro-c*,* or professional-level creative expertise. This level refers to creativity exhibited by individuals who have earned professional-level status in a discipline, but may not yet have transformed their field or made an eminent contribution. For scenario generation, Pro-c novelty is associated with scenario generation that produces scenarios at a level of quality, complexity, distinctiveness and unexpectedness as to offer significant value to a domain expert (i.e., instructor) possessing deep experience in the subject domain.

Pro-c level novelty is a target level for DEEPGEN, because of its promise to offer value to both instructors and advanced trainees alike. It calls for scenario generation functionalities that can operate upon virtually all aspects of a scenario within a given domain. In CFF training, this corresponds to modifying a CFF scenario's pre-mission briefing; augmenting the types, locations and behavior of friendly and adversary forces; inserting dynamic events at run-time (e.g., weather changes, communication failures); and altering the embedded scaffolding and assessment rules at play in scenarios. Notably, Pro-c level novelty need not exclusively pertain to complex orchestration of world states and triggered events. It is possible for two scenarios with identical units, terrain, locations, and scripted events to be distinguished from one another by augmenting the mission, or related context, faced by a trainee. When a prospective target enters an area, should the forward observer call for indirect fire, or let the target pass? How does the target's value compare to other possible targets? How much ammunition is available? What is the assigned objective for the forward observer? Questions such as these introduce meaningful forms of experiential novelty without requiring complex orchestration of events within a virtual simulation environment.

Given that Pro-c level systems are not necessarily targeted at the discovery of new military tactics or strategies, scenario generators like DEEPGEN can be designed to restrict the space of possible scenarios through deliberate authoring of the Scenario Adaptation Library, ensuring that all generated scenarios are both feasible and qualitatively different from one another. For example, several adaptable scenario elements listed in Table 1, such as changing how enemies react to being fired upon, are higher-level scenario adaptations that might typically be associated with the *Pro-c* level. This is in contrast to lower-level modifications, such as changes in weather or target type, that may be more likely to yield trivial variations between scenarios.

The next category, Little-c, refers to non-expert expressions of creativity, such as every-day problem solving or creative works. For scenario generation, we characterize this level of scenario generation in terms of creating "base-level" scenarios that enable trainees to reach basic proficiency in a domain. Little-c systems might only adjust a small number of features of a scenario, producing scenarios that are different by a small degree and do not require the level of domain knowledge or instructional expertise associated with professional-level scenario generation. In DEEPGEN, enacting scenario adaptations such as changing the weather, time of day, or type of target are all consistent with *Little-c* novelty. In some cases, this level of scenario generation may be preferable, because it is relatively inexpensive and efficient to setup and generate a large number of different, similarly structured scenarios that can be used for repeat practice of specific competencies.

The final level in the Four C model is mini-c. Mini-c describes creativity that is inherent to the learning process; it is expressed at an individual level while engaged in productive problem solving. This can be understood as a form of novelty that is experienced by novice learners as they begin to learn a new domain. For scenario generation, we define this level of novelty as consistent with generating "introductory" scenarios for a domain. These scenarios should be relatively simple from a complexity standpoint with novelty measured in relation to the concepts and competencies already achieved by a trainee.

Overall, the Four C model provides a useful framework for discussing and evaluating novelty in automated scenario generation systems. It provides a framework for formulating design requirements of these systems, and it points toward directions for evaluating the degree of novelty supported by scenario generators. Furthermore, it offers a useful ontology for describing how different characteristics of scenarios impact novelty within a given training domain.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

Automated scenario generation will serve a key role in the future of simulation-based training because of its significant potential for reducing the cost of creating novel scenarios and expanding access to high-quality simulation-based training. Data-driven approaches to automated scenario generation hold promise for enhancing trainee learning experiences by leveraging recent advances in reinforcement learning and interactive narrative technologies. We have presented an overview of DEEPGEN, a data-driven automated scenario generation framework, which formalizes the task in terms of enacting sequential adaptations to a canonical "parent" scenario in order to generate "child" scenarios that can be evaluated to assess learning outcomes. We have described an initial Scenario Adaptation Library that was developed for the domain of Call for Fire training. To better define and evaluate the degree of novelty embedded in scenarios generated by DEEPGEN, we have adopted the Four C Model, discussing how the model fits within the context of automated scenario generation and CFF training specifically.

In future work, we plan to expand DEEPGEN's Scenario Adaptation Library to capture a broader range of possible transformations to "parent" training scenarios, including sequential adaptations that can be performed dynamically over the course of a scenario. Further, it will be important to systematically investigate how instructors and learners interact with DEEPGEN, including the DEEPGEN Instructor Tool for configuring scenario generation, as well as generated scenarios for training a range of CFF skills. Finally, it will be critical to demonstrate how the DEEPGEN framework can be generalized to support additional domains. Integration with GIFT will be useful for enabling this line of investigation—automated scenario generation is particularly well-aligned with the Practice Quadrant in

GIFT's Engine for Management of Adaptive Pedagogy (EMAP)—setting the stage for expanding our understanding of how the Four C Model can be operationalized to measure and contextualize novelty in scenario generation across different domains.

ACKNOWLEDGMENTS

The research described herein has been sponsored by the U.S. Army Research Laboratory under cooperative agreement W911NF-18-2-0020. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

Borger, J. (2002). War game was fixed to ensure American victory, claims general. The Guardian, 21. Fairclough, C. (2004). Story games and the OPIATE system. University of Dublin, Trinity College.

Folsom-Kovarik, J.T., & Brawner, K. (2018). Automating Variation in Training Content for Domain-general Pedagogical Tailoring. In Proceedings of the Sixth Annual GIFT User Symposium (pp. 75-86). Orlando, FL: U.S. Army Research Laboratory.

Judson, J. (2018, September). 25 bloodless battles: Synthetic training will help prepare for current and future operations. Defense News, Retrieved from https://www.defensenews.com/smr/defense-news-conference/2018/09/05/25-bloodless-battles-synth etic-training-will-help-prepare-for-current-and-future-operations/

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four c model of creativity. Review of General Psychology, 13(1), 1-12.

Lee, S., Rowe, J. P., Mott, B. W., & Lester, J. C. (2014). A Supervised Learning Framework for Modeling Director Agent Strategies in Educational Interactive Narrative. IEEE Transactions on Computational Intelligence and AI in Games, 6(2), 203-215.

Luo, L., Yin, H., Cai, W., Zhong, J., & Lees, M. (2017). Design and evaluation of a data-driven scenario generation framework for game-based training. IEEE Transactions on Computational Intelligence and AI in Games, 9(3), 213-226.

Metz, C. (2016). In two moves, AlphaGo and Lee Sedol redefined the future. WIRED. com, 16.

Mott, B. W., & Lester, J. C. (2006). U-Director: A Decision-Theoretic Narrative Planning Architecture for Storytelling Environments. In Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (pp. 977-984). Hakodate, Japan: ACM.

Nelson, M. J., & Mateas, M. (2005, June). Search-Based Drama Management in the Interactive Fiction Anchorhead. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (pp. 99-104), Menlo Park, CA: AAAI.

Riedl, M. O., & Bulitko, V. (2012). Interactive narrative: An intelligent systems approach. AI Magazine, 34(1), 67.

Rowe, J., & Lester, J. (2015). Improving Student Problem Solving in Narrative Centered Learning Environments: A Modular Reinforcement Learning Framework. In Proceedings of the Seventeenth International Conference on Artificial Intelligence in Education (pp. 419-428). New York, NY: Springer.

Rowe, R., Smith, A., Pokorny, B., Mott, B., & Lester, J. (2018). Toward Automated Scenario Generation with Deep Reinforcement Learning in GIFT. In Proceedings of the Sixth Annual GIFT User Symposium (pp. 65-74). Orlando, FL: U.S. Army Research Laboratory.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. (2017). Mastering the game of go without human knowledge. Nature, 550(7676), 354-359.

Wang, P., Rowe, J. P., Min, W., Mott, B. W., & Lester, J. C. (2018) High-Fidelity Simulated Players for Interactive Narrative Planning. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (pp. 3884-3890). Stockholm, Sweden.

Young, R. M., Ware, S. G., Cassell, B. A., & Robertson, J. (2013). Plans and planning in narrative generation: a review of plan-based approaches to the generation of story, discourse and interactivity in narratives. Sprache und Datenverarbeitung, Special Issue on Formal and Computational Models of Narrative, 37(1-2), 41-64.

ABOUT THE AUTHORS

Dr. Jonathan Rowe *is a Research Scientist in the Center for Educational Informatics at North Carolina State University. He received the Ph.D. and M.S. degrees in Computer Science from North Carolina State University, and the B.S. degree in Computer Science from Lafayette College. His research is in the areas of artificial intelligence and human-computer interaction for advanced learning technologies, with an emphasis on game-based learning environments, intelligent tutoring systems, user modeling, educational data mining, and computational models of interactive narrative generation.*

Mr. Andy Smith *is a Research Scientist in the Center for Educational Informatics at North Carolina State University. He received his M.S. in Computer Science from North Carolina State University, and his B.S. degrees in Computer Science and Electrical and Computer Engineering from Duke University. Prior to graduate school Andy worked as an Underwater Robotics Engineer at SPAWAR SSC Pacific in San Diego, CA. His research is focused on the intersection of artificial intelligence and education, with emphasis on user modeling, game-based learning, and educational data mining.*

Dr. Randall Spain *is a Research Psychologist in the Center for Educational Informatics at North Carolina State University where he uses principles, theories, and methods of applied psychology (human factors, educational psychology, personnel psychology, experimental psychology, and psychometrics) to design and evaluate the impact of advanced training technologies on learning and performance. He has conducted training and human factors research for the Department of Defense and the Department of Homeland Security for the past 10 years with a focus on adaptive training, performance assessment and measurement, user modeling and human-automation interaction. Dr. Spain is a PhD graduate from Old Dominion University's Human Factors Psychology program and serves on the editorial board for Military Psychology.*

Dr. James Lester *is Distinguished Professor of Computer Science at North Carolina State University, where he is Director of the Center for Educational Informatics. His research centers on transforming education with technology-rich learning environments. With a focus on adaptive learning technologies, his research spans intelligent tutoring systems, game-based learning environments, affective computing, and tutorial dialogue. The adaptive learning environments he and his colleagues develop have been used by thousands of students in K-12 classrooms. He received his Ph.D. in Computer Science from the University of Texas at Austin in 1994. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI).*

THEME III:
LEARNER MODELING

Learner Modeling of Cognitive and Psychomotor Processes for Dismounted Battle Drills

Shitanshu Mishra[1], Gautam Biswas[1], Naveeduddin Mohammed[1], Benjamin S. Goldberg[2]
[1]Vanderbilt University - Institute for Software Integrated Systems,
[2]U.S. Army Research Laboratory

INTRODUCTION

Operations such as "Enter and Clear a Room" and "React to Direct Fire Contact" are essential dismounted battle drills (DBD) for urban warfare conducted by the armed forces. These operations require the soldiers to develop effective psychomotor and cognitive skills, and cognitive strategies along with the ability to work in teams. This paper discusses our initial research in developing intelligent tutors that support team training for DBDs in virtual and augmented reality environments.

As a first step toward developing tutors, we conduct an initial study of the "Enter and Clear Room (ECR)" DBD that relies heavily on team member psychomotor skills and cognitive skills, such as identifying and differentiate enemy combatants from noncombatants in the room, and providing cover for the other team members (Department of the Army, 2011). In addition to tactical skills, a squad also needs to develop strategic reasoning and decision making skills that are derived from situation awareness and planning to assure superior firepower inside and outside the building, determining the method of access into the building and rooms of interest, and for controlling the tempo of the operations (Holmquist & Goldberg, 2007). Since the operations are performed as a team, it is crucial that the trainees also acquire team skills in addition to the task skills (Sinatra et al., 2018). The need to combine individual psychomotor skills, cognitive and strategic processes, along with teamwork introduces a number of complexities in designing training scenarios and evaluating individual and team performance and effectiveness in these scenarios. The need to evaluate psychomotor, cognitive, strategic, and affective processes implies the need for multiple monitoring modalities, such as computer logs of individual and team performance, video analysis for analyzing psychomotor and cognitive skills, eye tracking for monitoring situation awareness, and physiological sensors to capture affect. Multi-modal data capture becomes even more critical when monitoring and analyzing complex teamwork.

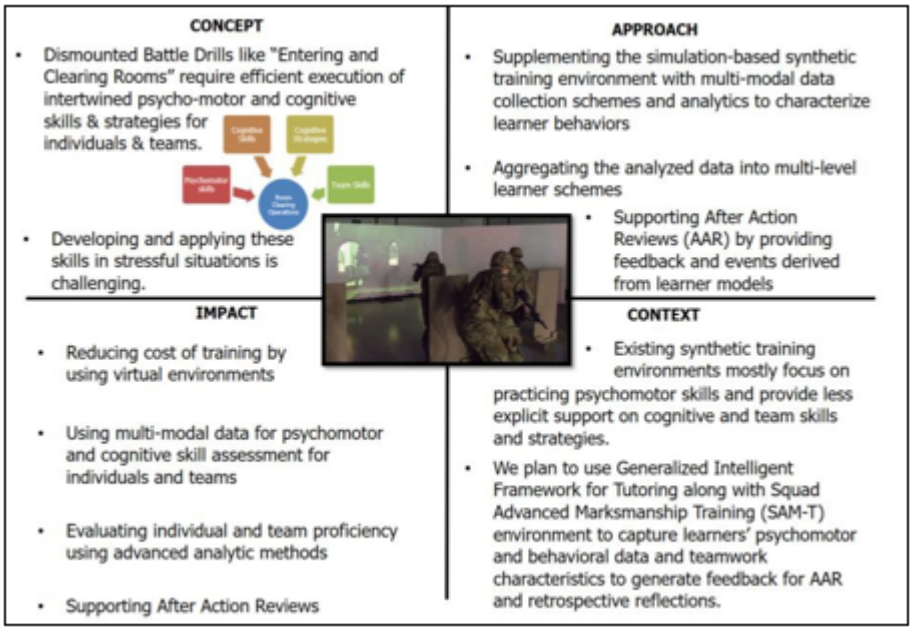


Figure 1. Quad chart showing the overall objective of the project

Figure 1 summarizes our overall approach to evaluating individual and team performance in synthetic training environments (STEs). In particular, we propose a number of performance and effectiveness metrics, propose, and corresponding measures for the Squad Advanced Marksmanship Trainer (SAM-T) implementation of ECR training operations for squads. We propose to compute these metrics within the Generalized Intelligent Framework for Tutoring (GIFT) framework (Sottilare et al., 2012), and develop integrated performance and effectiveness measures to support After Action Review by human instructors.

ENTER AND CLEAR ROOM (ECR) DOMAIN

"Enter and Clear Room" training scenarios are typically designed for a squad, i.e., a team of four soldiers. They represent a form of urban warfare, where enemy personnel have been located in a building that may also house noncombatants. The overall operations are initiated by securing the area around the building, and security forces are positioned in and around the building. A squad of four (sometimes two or three) are assigned to clear and secure a specific set of rooms, and the operation to clear each room begins on the order of the clearing team leader. It involves seizing control of the room by rapidly and tactically entering the room and neutralizing the enemy, while minimizing harm to the squad and the noncombatants. To accomplish this, the army divides up ECR missions into five major task segments Sinatra (2018): (1) Pre- pare to Enter, (2) Enter the Room, (3) Clear the Room, (4) Secure the Room, and (5) Completion and move on to next assigned operation. In this paper we focus on the segment of "Clear the room". Accompanying these tasks are well-defined *rules of engagement* (ROE). In this paper, we focus on task segments (2) and (3), i.e., ECR, which involves entering a room quickly and stealthily, moving immediately to *points of domination* (POD) while eliminating enemy combatants with superior fire power, and once clear seize control of the room. Figure 2 illustrates the tasks steps related to ECR. They are summarized below:

STEP 1. The squad in tight formation readies to enter the room, checks for booby traps on the door, and on a signal, usually a pat or an arm squeeze from the team lead (usually at the second position), is passed on to the first, starts the entry process (this may involve kicking down the door).

STEP 2. The first two Soldiers enter the room almost simultaneously. (Figure 2a). The first Soldier enters the room and moves left or right along the *path of least resistance* (typically the wall) to one of two corners. The soldier enters firing aimed bursts into his sectors engaging all threats or hostile targets to cover his entry. He assumes a POD facing into the room.

STEP 3. The second Soldier enters the room immediately after the first Soldier. He moves in the opposite direction of the first Soldier to his point of domination, also firing aimed bursts to engage and eliminate all threats in his sector.

STEP 4. The third Soldier moves in the opposite direction of the second Soldier while scanning and clearing his segment of the room. In some situations, the third soldier is assigned to cover threats from the top, i.e., the ceiling or gaps that may exist in the ceiling. (Figure 2b)

STEP 5. The fourth Soldier moves opposite of the third Soldier to a position that dominates his sector, also scanning and clearing his assigned region. (Figure 2c)

STEP 6. All Soldiers are positioned at their PODs as they continue to scan their sectors and engage enemy combatants with precision aimed fire, while avoiding injury to the noncombatants.

STEP 7. The team assesses if the room is neutralized. The team leader announces (or sends message) to the squad leader when the room is "CLEAR."

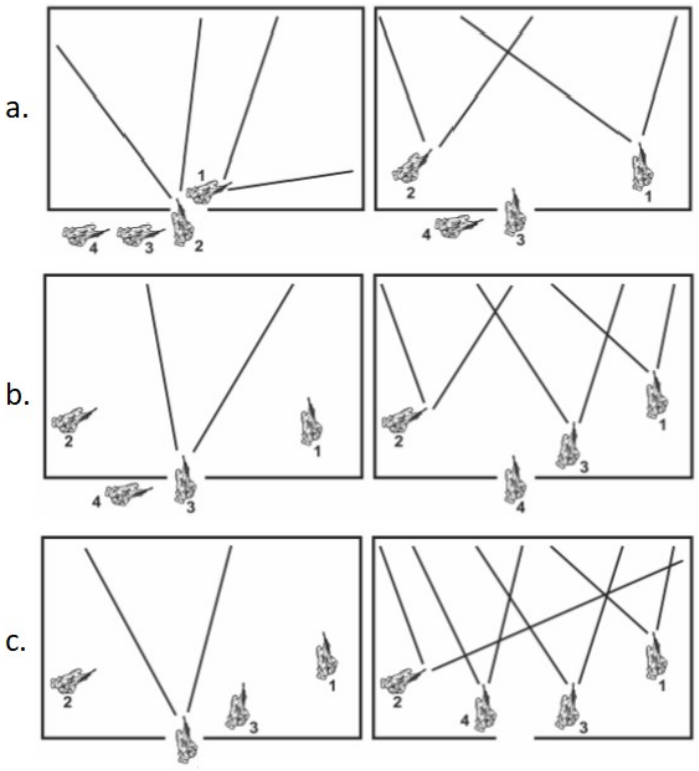


Figure 2. Clear a room: a) First two Soldiers enter almost simultaneously; b) Third soldiers enters; c) Fourth soldier enters

Expert interviews made us aware that a lot of rapid assessments and decisions need to be made by the soldiers from the start to the end of the clearing operation. In other words, proficiency in the clear operation requires superior *muscle memory* (Scales, 2013). For example, at the start of entering the room, the position of the door hinges may influence the direction of the movement of the first and the second soldier. If the door hinge is on the left (right) then the first soldier turns to the right (left) side. The second soldier goes in the direction opposite to the first. A second rapid decision a soldier needs to make is whether a person in the room is a combatant or not. A number of factors, e.g., possession of weapon and whether the person kneels when commanded, influence such decisions. We discuss psychomotor and cognitive skills and strategies for ECR operations in greater detail below.

ECR TRAINING ENVIRONMENT: SAM-T

The Squad Advanced Marksmanship Trainer (SAM-T) is a Training as a Service (TaaS) solution designed to enable army readiness and bridge the dismounted virtual collective training capability gap pending fielding of the Soldier/Squad Virtual Trainer in 2021. SAM-T (Figure 3) is an augmentation (not a replacement) of Engagement Skills Trainer (EST) II, which was designed to simulate live weapon training events that directly support individual and crew-served weapons qualification, including individual marksmanship, small unit collective and judgmental escalation-of-force exercises in a controlled environment [https://asc.army.mil/web/portfolio-item/engagement-skills-trainer-est/]. SAM-T is intended to improve and accelerate Soldier and Squad close combat skills, and task acquisition by providing the realistic repetitions in diverse complex operational environments necessary to increase readiness.

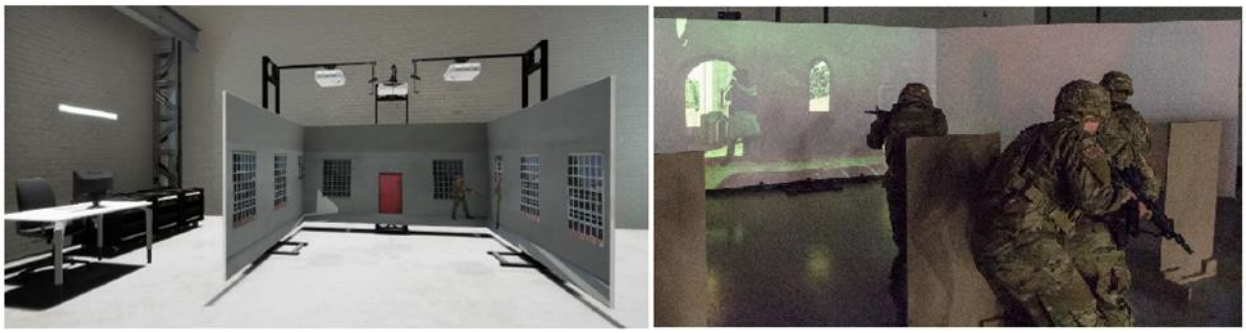


Figure 3. U-Shape System Configuration for SAM-T (Pargett, 2019)

The expected capability listing for SAM-T include: (i) Weapon Skill Development: An immersive individual, crew, and collective weapon skill development training capability; (ii) Use of Force (UoF): Use of Soldier cognitive functions to include rapid decision making and target acquisition in stressful scenarios; (iii) Battle Drill Training: This dismounted maneuver requires the capability to conduct collective battle drills and tasks. Basic collective task training that SAM-T provides include: (i) Enter and clear a room; (ii) React to direct fire contact while dismounted; (iii) Employ hand grenades; and (iv) Use visual signaling techniques. The scope of our project is limited to ECR operations. Training in the SAM-T is customizable as per individual/team readiness and requirements. Variation and combination of stressors address multiple customizations and scenarios. Three variations are included: (1) change in physical layout (e.g., obstacle, training area, etc.); (2) change in physical parameters (e.g., target distances, target movement, and target appearance, etc.); and (3) Human factors (Callisthenic tasks, Cognitive tasks, Personal Equipment variations, etc.)

Proposed Design and Analysis

We propose to integrate SAM-T with GIFT and action review modules (Figure 4). The user behavior logger attached to the SAM-T environment includes multiple sensors with different modalities. We anticipate to have data collected from Integrated visual augmentation system (IVAS), audio communication data, data from head-mounted eye trackers, data related to gun, for example: x-y-z coordinate locations of the gun from the screens, weapon trigger events, weapon states, bio harness sensors, and video observations. Trainees, in teams of 3–5 would enter the virtual room (the SAM-T environment) to perform and practice ECR operations. The user behavior logger module will log the soldier movement data in video and motion tracking form. The GIFT module will be designed to analyze the multimodal data and generate feedback for the *After Action Review* (AAR) module. Overall, GIFT module will primarily perform four functionalities: (i) Detection of task and team skills attributes from the data; (ii) Evaluation of individual and team performance and identification of deficient skills and strategies at individual and team levels; (iii) Learner modelling to aggregate and keep track of learner and team performances for various skills and strategies over multiple practice scenarios; (iv) Feedback generation for individual and for team based on reports generated from the learner model. The AAR module, with or without the presence of the human moderator (coach) will provide formative feedback and replays of the trainee's behavior to help them to perform retrospective reflections and regulate their performance in subsequent practice iterations. The action review module can have two parts: (i) mid action reviews, where the feedback and reviews are given in between the practice session without any moderation of the external human agent (i.e., coach). Whereas, in the case of after action reviews, the action review module can be moderated by the human coach to help the trainees reflect on their performance.

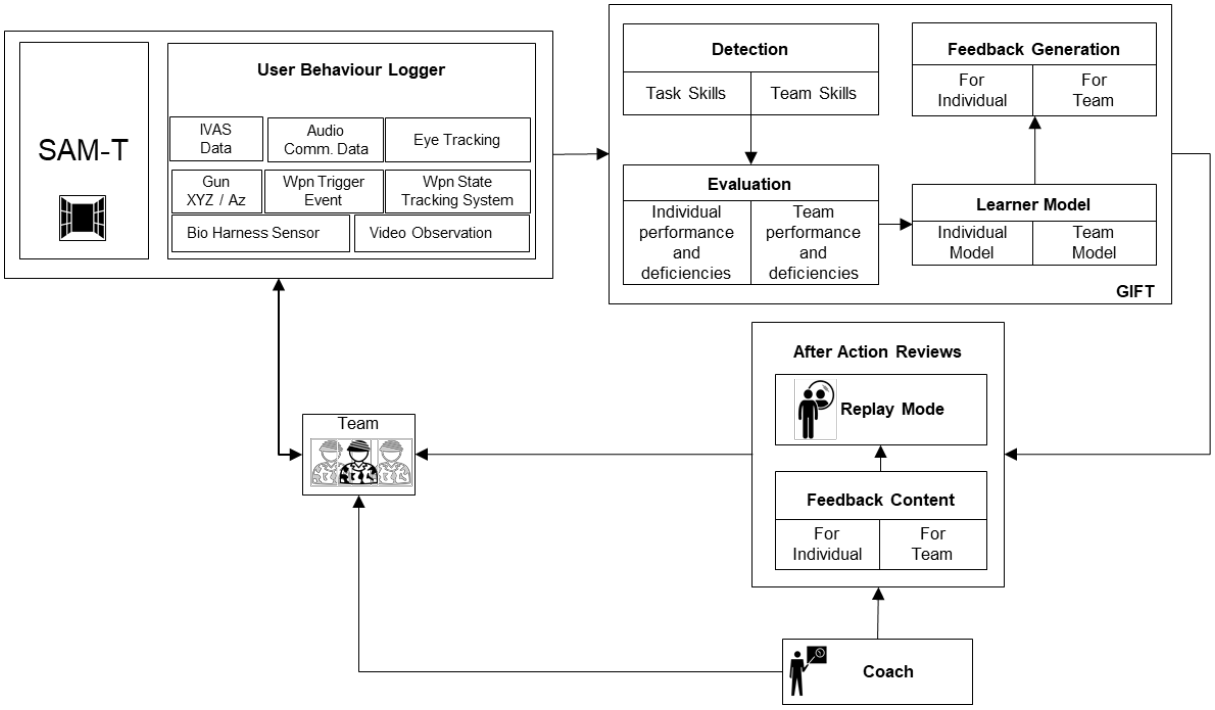


Figure 4. Proposed integration of SAM-T with GIFT and Action Review modules

Tracking Learner and Team Proficiencies: Method

Analyzing team tasks and defining individual task and team skills along with quantitative measures of performance and effectiveness (MOP and MOE) represent the initial steps towards developing team tutors. In addition to referring to army training manuals (especially the Army Training Registry and Central Army Registry (CAR)), we also conducted informal interviews with Subject Matter Experts (SMEs) to gain an overall understanding of the ECR processes and steps at the psychomotor, cognitive, and strategic levels. Our expert interviews started with a broad question: "What are the key characteristics and key steps in the ECR process? Once we had an overall understanding of the ECR process and steps (the important pre-, during-, and post-activities), we dove deeper into the psychomotor, tactical, strategic, cognitive and team skills that the soldiers needed to exhibit for successful ECR operations?" In addition, as the interviews progressed, we asked a lot of "what if" questions, mainly to gain an understanding of how standard operating procedures (SOPs) might deviate, when unusual situations were encountered. Inductive thematic analysis of the interview transcripts was performed to extract categories of psychomotor and cognitive skills and strategies (Guest, MacQueen, & Namey, 2011). This form of task analysis produced the list of psycho- motor skills and methods for measuring these skills as has been described in Tables 1 and 2.

Measures

The learning environment will compute a number of individual and team performance measures as trainees practice ECR scenarios. These performance measures, along with knowledge of trainees' actions in the environment have to be logged, such that they can further be used for analyzing learners' proficiencies related to the ECR domain. Table 1 shows a list of performance measures for proficiencies relevant to the "clearing room" segment of the ECR operation. The second and third column of the table provide preliminary definitions of how these measures are computed and the sensing modality that provides the information for computing the measures, respectively. The list of measures shown in the table includes proficiencies that are required: (1) 'just' before entering the room through the door; (2) executing the clear operations in the room; and (3) securing and executing the room after the clear operations are completed. We have listed measures of performance and effectiveness that are primarily relevant to the "clearing" segment. We plan to refine and expand these measures through further consultations with our subject

matter experts, collecting and analyzing video data of soldier movements during the move, and learning more about the variations to the standard clear scenarios and how the standard operating procedures are modified to adapt to these scenarios.

Table 1. Performance measures in ECR operation

Performance Measures		Measure Values	Data Source
M1	Soldiers (Team of 4) line up at the door	M1 = 0, if trainee is not on the door location = 1, if trainee is on the door location	Video Observation
M2	Keeping eyes in assigned regions	M2 = [0,1], normalized angular deviation	Eye Tracker
M3	Speed of Entry	M3 = 0, if time difference is > threshold 1, if time difference is < threshold	Video Observation
M4	Concealed presence	M4 = 0, if needless talking or exposing the tip of a rifle across an open doorway = 1, otherwise	Video Observation
M5	Signal to commence operation (Leader)	M5 = 0, failed to deliver signal = 1, successful, used SOP for communication during entering (either Triceps squeeze, Shoulder squeeze, or Muzzle dip)	Video Observation
M6	Enter the room in the correct direction	M6 = 0, if direction is opposite to the direction of previous trainees = 1, otherwise	Video Observation
M7	Moving along the wall	M10 = [0, 1], normalized count of number of conditions satisfied from below: 1. Trainee continues moving while clearing 2. Stops if reached corner or to a POD 3. Continues moving along one wall after reaching corner 4. Speed of movement is in the range where they can move while accurately engaging any targets 5. Did not over-penetrate into the room	Video Observation
M8	Identifying adversaries	M7 = $(M7_{gaze} + M7_{gun})/2$ $M7_{gaze}$ = [0,1], normalized angular deviation between gaze azimuth and the position of the adversary, $M7_{gun}$ = [0,1], normalized angular deviation between gun azimuth and the position of the adversary	Eye Tracker, Gun Azimuth
M9	Eliminating nearest threat	M8 = [0, 1], normalized count of number of conditions satisfied from below: 1. Threat is dealt before clearing the near corner 2. Either a minimum of one round shot if adversary is armed or a well-placed arm check if unarmed 3. Not stuck into firefight 4. Not stuck into addressing a potential threat deep in the room	Video Observation, Gun Azimuth, Gun Trigger
M10	Watching the near corner	M9 = [0, 1], how quick the trainee has scanned/ cleared the near corner, 0: Fixated for too long, 1: Fixated for minimum required time	Eye Tracker
M11	Collapsing a sector	M11 = 1, if scan overlaps teammates sector of fire by a threshold = 0, if scan overlaps teammates sector of fire by more than a threshold = 0, if scan doesn't overlaps teammates sector of fire	Gun Azimuth, Eye Tracker

M12	In-operation Communication	M12 = [0, 1], normalized count of number of conditions satisfied from below: 1. Communicated if any movement outside of the tactic 2. Communicated danger areas if found 3. Communicated any intent to move deeper into the room for a search	Audio Comm. Data, Video Observation
M13	Non-combatant casualty	M13 = -1, if gun triggered and gun azimuth aligns with the position of a civilian or non-combatant = 0, otherwise	Gun Azimuth, Gun Trigger
M14	Marksmanship	M14 = normalized marksmanship score with respect to the current target	Marksmanship performance
M15	Reporting exit	M15 = 1, if used SOP for communication during exiting (i.e. Sent "Clear" message) = 0, otherwise	Audio Comm. Data
…	…	….	….

Table 2 shows how different performance measures can inform overall individual and team performances at each step of the "clearing room" segment of ECR operation. (These steps correspond to the seven steps during the room clearing task discussed at the beginning of this section.) To compute the overall performance of an individual trainee, performance values corresponding to that trainee across all the steps have to be aggregated. Whereas, to compute the team performance for any specific step, performance values corresponding to that step across all the trainees have to be aggregated. The overall team performance can be the aggregation of the team performances at individual steps.

Table 2. Aggregating individual and team performances

Trainee	Step1 (S1)	Step2 (S2)	Step3 (S3)	Step4 (S4)	Step5 (S5)	Step6 (S6)	Step7 (S7)	Overall
T1	f(M1, M2, M3, M4)	f(M6, M7, M8, M9, M10, M13, M14)	f(M7, M8, M9, M10 M12, M13, M14)	f(M7, M8, M9, M10, M12, M13, M14)	f(M7, M8, M9, M10, M12, M13, M14)	f(M7, M8, M9, M10, M11, M12, M13, M14)	f(M15)	h(S1:S7)
T2	f(M1, M3, M4, M5)	f(M4)	f(M7, M8, M9, M10, M13, M14)	f(M7, M8, M9, M10, M12, M13, M14)	f(M7, M8, M9, M10, M12, M13, M14)	f(M7, M9, M10, M11, M12, M13, M14)	f(M15)	h(S1:S7)
T3	f(M1, M4)	f(M4)	f(M4)	f(M6, M7, M8, M13, M14)	f(M7, M8,, M9, M10, M12, M13, M14)	f(M7, M9, M10, M11, M12, M13, M14)	f(M15)	h(S1:S7)
T4	f(M1, M4)	f(M4)	f(M4)	f(M4)	f(M6, M7, M8, M12, M13, M14)	f(M7, M9, M10, M11, M12, M13, M14)	f(M15)	h(S1:S7)
Team	g(T1:T4)	g(T1:T4)	g(T1:T4)	g(T1:T4)	g(T1:T4)	g(T1:T4)	g(T1:T4)	h(S1:S7)

LEARNER MODELING

In past work, we have developed hierarchical learner modeling schemes that involve open-ended learning in K-12 environments (Rajendran et al., 2017, Kinnebrew et al., 2017) and problem solving involving complex decision making tasks (Biswas et al., 2019, in review). The learner modeling scheme is meant to analyze and represent trainee's proficiencies in complex decision-making scenarios. In the current project, our overall goals are to extend this multi-level learner modeling approach to Battle Drill domains. One significant addition to this framework will

be the addition of a psychomotor task modeling layer to the existing three-layer model that we have developed in previous work.

At each step, the learner model will use trainee's performance on specified goals (and sub-goals) and tasks (the measures were discussed in the last section) to update the different levels of the hierarchy (Figure 5a). The cognitive (and psychomotor) skills layer of our learner model will derive information of a trainee's interactions from behavior logs generated by the Behavior Logger module (Figure 4). These behavior logs will serve as an assessment of leaner's ability to execute domain-specific knowledge and skills. For example, during the clearing operation, the trainee has to tactically maneuver inside the room that requires him to move along the wall. The learning environment will keep track of whether the trainee moves well along the wall (M7 in Table 1) while he is inside the room or not. If the trainee does not conform to the standards of moving along the wall (M7) the learner-modeling framework attributes this to trainee's lack of 'tactical maneuvering' skill. The next layer, the cognitive strategies level, will involve conditional knowledge about how to combine situation-specific information and cognitive/ psychomotor skills to accomplish higher level tasks and goals (Kinnebrew et al., 2017). The conditional aspect of cognitive strategies involves under- standing when a strategy is most effective, especially when there are multiple potential courses of action available to the trainee. The top layer will use tracked changes in trainee's performance on specified goals and strategies, and observable interactions with the After Action Review module to capture the trainee's proficiency in metacognitive processes.

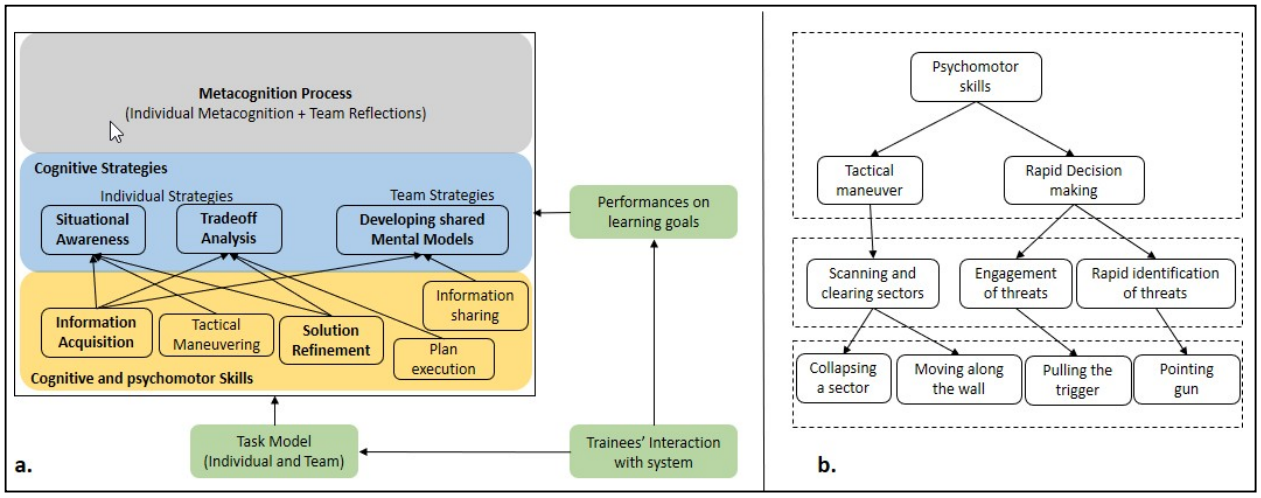


Figure 5. a) Three Tier Hierarchical Learner Model; b) A part of task model for psychomotor skills

Each practice sessions of ECR are very short (5 secs to 60 secs for inside-room operation). For such a complex and rapid decision-making task, trainees continually apply skills and strategies to make decisions and conduct operations within and across practice iterations. Hence, we accumulate trainees' proficiencies in skills and strategies as a function of time (practice iterations in SAM-T). We will use the performance metrics 'competence' and 'trend' (Biswas et al., 2019, in review) to measure the evolution of learners' proficiencies in skills and strategies in SAM-T. Competence (Ct) captures the learner's accumulated proficiency on a skill or strategy, while the trend value (T) for a specific skill or strategy represents a local measure of how the learners' competence evolves with practice iterations. As defined by Biswas et al. (2019, in review) competence (Ct) at any time iteration t is defined as the sum of the learner's competence in that skill (strategy) at iteration, t - 1, and the value representing an aggregation of performances on all observable actions relevant to any skills or strategy at the iteration t, i.e.,

$$C_t = C_{t-1} + f(\text{performance on observable actions})$$

The trend value (T) can be computed as a function of the change in competence over the last two iterations, defined as:

$$T = g((C_t - C_{t-1}); (C_{t-1} - C_{t-2}))$$, where g can be defined by the system designer.

To compute proficiency in a strategy, we combine trainees' performance in a related sub-goal and related skill(s). When learner's competence in executing a skill and performance in achieving a related sub-goal, both are positive, that implies a positive application of the related strategy, and, therefore, we increment the proficiency in the strategy positively. Whereas, a non-positive performance in sub-goal indicates inefficiency in the application of related strategy, irrespective of whether the competence in executing the related skills is positive or not. However, in a weird case, when sub-goal performance is positive but skill competence is negative, the effectiveness of applying the strategy is hard to detect.

An essential precursor to authoring any team tutor is the analysis of team tasks and defining task skills and team skills and thereby creating a task model that contains a hierarchy of learning environment (LE) – general skills, LE-specific tasks and observable behaviors in LE. The tasks model informs the lowest layer (skills layer) of the learner model with the observable behaviors linked with the skills. Figure 5b presents an example of part of the task model and presents a subset of psychomotor skills, extracted during our initial task analysis. Figure 5b elaborates two skills (i) tactical maneuvering; and (ii) rapid decision making. Ac- cording to Figure 5b, tactical maneuvering requires collapsing a sector (M11) and moving along the walls (M7) in the room. Performance measures M11 and M7 in Table 1 can be used to measure trainee's proficiency in tactical maneuvering. Similarly, Rapid decision making involves instantaneous decision making and responding by pointing the gun towards or shooting at inhabitants, once they are identified as enemy combatants. Similar to the part of the task model corresponding to the psychomotor skills, the task analysis also provided sets of relevant tasks that require cognitive skills and team skills for the phases before, during and after clearing room.

CONCLUSIONS

In this paper, we presented our initial work towards the creation of an intelligent learning environment that supports the training of army personnel on the skills and strategies necessary for successfully conducting the ECR operations. Preliminary takeaways from referred documents and subject matter experts have revealed that speed (time), accuracy, and marksmanship are key factors for a successful ECR operation. Trainees have to develop muscle memory, which can only be acquired through iterations of practice. Skills involving rapid decision making, for example, identification of combatants and non-combatants requires iterations of practice with variations in the room inhabitants across multiple practice iterations. In addition to the individual task skills, team skills are also needed to ensure that the trainees efficiently follow standards of procedures, adhere to the rules of engagements, communicate, and avoid disastrous fratricide.

It should be noted that the exact protocols of entering and clearing the room may not be replicated all the times. For example, the first soldier may not move along the wall if there is furniture in his path, and, therefore, he may have to improvise the path. Similarly, based on the room configurations (objects and inhabitants in the room) the soldier may have to choose different domination points, other than room corners. Therefore, default trajectories and the rules of engagement discussed in this paper do not apply universally, but have to be modified to accommodate characteristics of the actual scenario. This makes analyzing user performance and effectiveness in the ECR domain more challenging. The need to evaluate both individual and team skills and performance adds to the challenge. As we proceed, more expert interviews, literature synthesis and observations of trainees performing the ECR operations are needed to enrich the task model and the computation of measures of performance and effectiveness (MOP and MOE) to update the task model. Capturing multimodal learner data as described in this paper, would provide the framework for accurately measuring learners' performances in such complex scenarios that require keeping track of psychomotor, cognitive and team skills, and infer their cognitive strategies. We also look forward to further enhance the team assessment by incorporating the Squad Performance Model to measure the team Lethality.

REFERENCES

Biswas, G., Rajendran, R., Mohammed, N., Goldberg, B.S., Sottilare, R.A., Brawner, K., and Hoffman, M. (2019, in review). "Multilevel Learner Modeling in Training Environments for Complex Decision Making," IEEE Transactions on Learning Technologies.

Guest, G., MacQueen, K. M., & Namey, E. E. (2011). Applied thematic analysis. Sage Publications.

Fletcher, J. D., & Sottilare, R. A. (2018). Shared mental models in support of adaptive instruction for teams using the GIFT tutoring architecture. International Journal of Artificial Intelligence in Education, 1-21.

Holmquist, J. P., & Goldberg, S. L. (2007). Dynamic Situations: The Soldier's Situation Awareness. University of Central Florida Orlando.

Kinnebrew, J. S., Segedy, J. R., & Biswas, G. (2017). Integrating model-driven and data-driven techniques for ana- lyzing learning behaviors in open-ended learning environments. IEEE Transactions on Learning Technolo- gies, 10(2), 140-153.

Pargett, M. (2019, March 25). Squad tactics tested on new virtual marksmanship trainer. Retrieved April 2, 2019, from https://www.army.mil/article/219231/squad_tactics_tested_on_new_virtual_marksmanship_trainer

Rajendran, R., Mohammed, N., Biswas, G., Goldberg, B. S., & Sottilare, R. A. (2017). Multi-level User Modeling in GIFT to Support Complex Learning Tasks. In Proceedings of the 5th Annual Generalized Intelligent Frame- work for Tutoring (GIFT) Users Symposium (GIFTSym5).

Randel, J. M., Pugh, H. L., & Reed, S. K. (1996). Differences in expert and novice situation awareness in naturalistic decision making. International Journal of Human-Computer Studies, 45(5), 579-597.

Scales, B. (2013). Virtual immersion training: bloodless battles for small-unit readiness. The Magazine of the Asso- ciation of the United States Army, 24-27.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. Research in science education, 36(1-2), 111-139.

Sinatra, A. M., Kim, J. W.,2, Johnston, J., Sottilare, R. A. (2018). Assessment of Team Performance in Psychomotor Domains. Design Recommendations for Intelligent Tutoring Systems: Volume 6. US Army Research Labor- atory.

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). The generalized intelligent framework for tutoring (GIFT). https://gifttutoring.org/projects/gift/wiki/Overview

U.S. Department of the Army. (2011, June 10). Army Tactics, Techniques and Procedures – ATTP 3-06.11.

ABOUT THE AUTHORS

Dr. Shitanshu Mishra *is a Postdoctoral Researcher at the Institute of Software Integrated Systems at Vanderbilt University.*

Dr. Gautam Biswas *is a Professor of Computer Science, Computer Engineering, and Engineering Management in the EECS Department and a Senior Research Scientist at the Institute for Software Integrated Systems (ISIS) at Vanderbilt University.*

Naveeduddin Mohammed *is a Research Engineer at the Institute of Software Integrated Systems at Vanderbilt University.*

Dr. Benjamin Goldberg *is an adaptive training scientist at the Army Research Laboratory's SFC Paul Ray Smith Simulation & Training Technology Center. He leads research focused on instructional management within ARL's Learning in Intelligent Tutoring Environments (LITE) Lab and is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT).*

Towards Deeper Integration of Intelligent Tutoring Systems: One-way Student Model Sharing between GIFT and CTAT

Vincent Aleven[1], Jonathan Sewall[1], Juan Miguel Andres[2], Octav Popescu[1], Robert Sottilare[3], Rodney Long[3], Ryan Baker[2]
Carnegie Mellon University[1], University of Pennsylvania[2], US Army Research Laboratory[3]

ABSTRACT

There are strong potential benefits to be had by integrating intelligent tutoring systems (ITSs) with each other, but few instances of successful integration are known in the literature. Given the central role that student models play in ITS architectures, and given that different ITS platforms tend to have their own student models, a key challenge is exchanging and mapping student models. Our project focuses on integrating GIFT and CTAT, both widely used ITS authoring and delivery environments. The specific goal of our project is to create, as a proof-of-concept, adaptive capabilities for an edX MOOC, using GIFT and CTAT within edX. We have created an initial version of this integration, in which GIFT supports the outer loop and simple interactive activities, while CTAT (under GIFT outer loop control) supports more complex problem-solving activities. As a student works with a CTAT tutor, whenever CTAT updates its own student model, the updates are sent also to GIFT, so that GIFT's outer loop can take advantage of a complete and up-to-date view of the student's knowledge as it selects appropriate remedial activities. The main student model elements are mapped in a 1:1 manner between CTAT's and GIFT's student model, even if the general problem of creating such mappings is hard. This one-way student model sharing is achieved with an extended use of the LTI standard. The main contribution is a proof-of-concept demonstration of ITS integration, limited in a number of ways (e.g., for the time being, the student model is communicated in one direction only), but exciting in its possibilities for joining the ITS functionality of different ITS platforms.

INTRODUCTION

Intelligent tutoring systems can be authored, increasingly, with efficient and easy-to-learn authoring tools, such as the Generalized Intelligent Framework for Tutoring (GIFT), (Brawner, 2015; Goldberg & Hoffman, 2015; Goldberg, Hoffman, & Tarr, 2015; Sottilare, 2012), the Cognitive Tutor Authoring Tools (CTAT) (e.g., Aleven et al., 2016), and others (Cai, Graesser, & Hu, 2015; Mitrovic et al., 2009; Razzaq et al., 2009). Although different ITSs tend to share a core of tutoring behaviors (VanLehn, 2006; 2016), they often have complementary strengths and focuses. As noted in Baker (2016), the challenge of developing a single form of adaptivity is often sufficiently high that some ITSs focus on just one form of adaptivity apiece; other ITS include multiple forms of adaptivity, but not always the same forms (Aleven, McLaughlin, Glenn, & Koedinger, 2017). GIFT, for example, offers an adaptive outer loop that covers a wider range of pedagogy than CTAT; it also offers tools for easy authoring of questions to test recall of concepts and APIs for integrating sensors and training applications. CTAT, on the other hand, offers possibilities for crafting highly adaptive step loops, responsive to students' strategies and errors, and offers an adaptive outer loop that supports cognitive mastery.

A promising approach to building effective, innovative, adaptive learning technologies would therefore be to bring together ITS systems to leverage the strengths found in each (integrating GIFT and CTAT, for instance). Potential advantages could be, speculatively, that more adaptive tutoring systems could be authored more easily, that systems with more sophisticated pedagogical approaches could be authored, and that the choice of pedagogy could be better matched to the instructional goals.

ITS interoperability has long been viewed as desirable (Brusilovsky, 1995), but has proven elusive, now forming the basis of one of the BLAP prizes in Learning Analytics (Baker, under review). A small number of interesting instances exist (Aleven & Rosé, 2004; Koedinger, Suthers, & Forbus, 1998), but the main ITS platforms are still separate.

Given the central role that student models play in ITS architectures (Bull & Kay, 2016), sharing or mapping student models should be a focal point in the integration of ITSs: if tools could share their student models, then tutors created with these tools should be able to make better adaptive decisions from the richer, more complete information available (Aleven et al., 2017; Woolf, 2009), and make better decisions sooner when a student starts in a new platform (Sosnovsky et al., 2007). However, the student models used in different ITS platforms tend to differ in the types of student characteristics they assess, the ontologies that they use to represent these characteristics, their methods for updating the model, and the data they require. The many differences make integration of student models, at least as a general problem, quite a daunting prospect. In the current project, we explore a small but interesting instance of this challenge.

Specifically, our project focuses on creating a MOOC, within the edX platform, that is adaptive to students' knowledge growth in ways that edX courses are not. We do so by integrating GIFT and CTAT with each other, and embedding them together within edX, so that GIFT's and CTAT's combined adaptive tutoring functionality is available in the MOOC. We carry out this integration and demonstrate its feasibility in the context of the edX MOOC "Big Data and Education" (BDEMOOC), created and taught by the last author. In the current paper, we focus on the GIFT/CTAT integration, as we reported on the integration into edX in prior publications (Aleven, Baker, et al., 2017). Also in prior work, we made it possible for GIFT to invoke CTAT tutors in a manner adaptive to a student's knowledge growth as assessed by GIFT (Aleven et al., 2018).

We now extend this work so the CTAT tutor can send its up-to-date student model to GIFT. This model captures a student's mastery of knowledge components targeted in the instruction (Aleven et al., 2016; Aleven & Koedinger, 2013). We enabled GIFT to map CTAT's student model onto its own student model, which (among other things) captures similar knowledge components. This way, GIFT's outer loop has up-to-date information about a student's skill level on which to base the adaptive selection of learning activities. Although one could envision other ways of combining GIFT and CTAT, this particular way plays to the strengths of both tools, as discussed in more detail below.

In the current paper, we address the following questions: What leverage is there in enabling CTAT to communicate its student model to GIFT? What adaptive tutoring behaviors might now be easier to author than before? How can the two student models be mapped to each other? How can their integration be accomplished technically? What are the limitations of this means of integrating student models, and how might they be addressed in future work?

ADVANTAGES OF INTEGRATION: TARGETED TUTORING BEHAVIORS

In this section, we describe the student experience that we implemented within the BDEMOOC as a proof-of-concept demonstration of the new GIFT/CTAT integration. One could envision more complex forms of adaptive instruction based on this integration, but we wanted to start simple. Specifically, we added a new pattern of adaptive instruction that includes examples and learn-by-doing activities for week 1 of the 8-week BDEMOOC, implemented as a short GIFT course embedded within the overall edX course. The course used as its outer loop GIFT'S Engine For Management of Adaptive Pedagogy (EMAP), which implements Merrill's component display theory (CDT) quadrants (Goldberg et al., 2015). Generally speaking, in EMAP a student first enters the optional Rules quadrant to receive direct explanation of the concepts (e.g., in a video lecture), then proceeds to the Examples quadrant to see instances of application of the concepts. Next, in the Recall quadrant, the student answers questions associated with individual concepts. The optional Practice quadrant specifies activities by which the student can learn or demonstrate skill with applying the concepts. The Remediation "quadrant" (not explicit in the original CDT) provides concept-specific materials for review if the student's Recall or Practice performance does not meet expectations. An author defines the quadrants by configuring GIFT's Adaptive Courseflow object with the concepts and materials to be presented.

In our edX MOOC, week 1 includes 6 short GIFT courses, each with an Adaptive Courseflow object having a video lecture (Rules content) explaining concepts and techniques in educational data mining (the subject of the course)

and providing slides with examples and recall questions on these concepts. We split the material into individual GIFT courses in order to provide questions after each lecture, instead of after all 6 lectures, and to let students use edX to navigate to individual lectures atwill.

New in the 2019 edition of our MOOC is a 7th GIFT course at the end of week 1, with an Adaptive Courseflow object configured as shown in Figure 1. This course covers concepts and skills related to Decision Trees and *k*NN (for *k*-Nearest Neighbor). The course's principal purpose is to permit GIFT to provide adaptive practice with a CTAT tutor, and then depending on the success of the student's learning in the CTAT tutor, to present opportunities for remedial studying of examples, and remedial additional practice with a second tutor. To this end, CTAT communicates its student model to GIFT, to summarize the state of student learning resulting from the tutor activity.

Although an Adaptive Courseflow typically starts in the Rules quadrant, the new course's Adaptive Courseflow object omits explicit Rule content, to avoid repeating material from the video lectures earlier in the week. Its Examples quadrant (top left in Figure 1) provides detailed Powerpoint slides illustrating the application of the two algorithms (Decision Trees and *k*NN). The Recall quadrant (top right in Figure

1) has click-through screens instead of questions, again to avoid redundancy with the questions asked earlier. The Practice quadrant (bottom right in Figure 1) offers two CTAT tutors as practice applications: the primary tutor, presented first, covers both algorithms. If the student's skill level after exiting the primary tutor is still Novice on either algorithm, then the remediation quadrant (bottom left, Figure 1) lets the student review the example slides for just that algorithm. After remediation, the student will do a secondary tutor that covers both algorithms. We would have preferred to have two separate tutors for remedial practice, one for each algorithm, but GIFT disqualifies from remedial use any Practice application that fails to cover *all* Practice quadrant concepts, even those already mastered. Even so, the student still receives adaptive content due to the integration of CTAT tutors in the Remediation Quadrant.

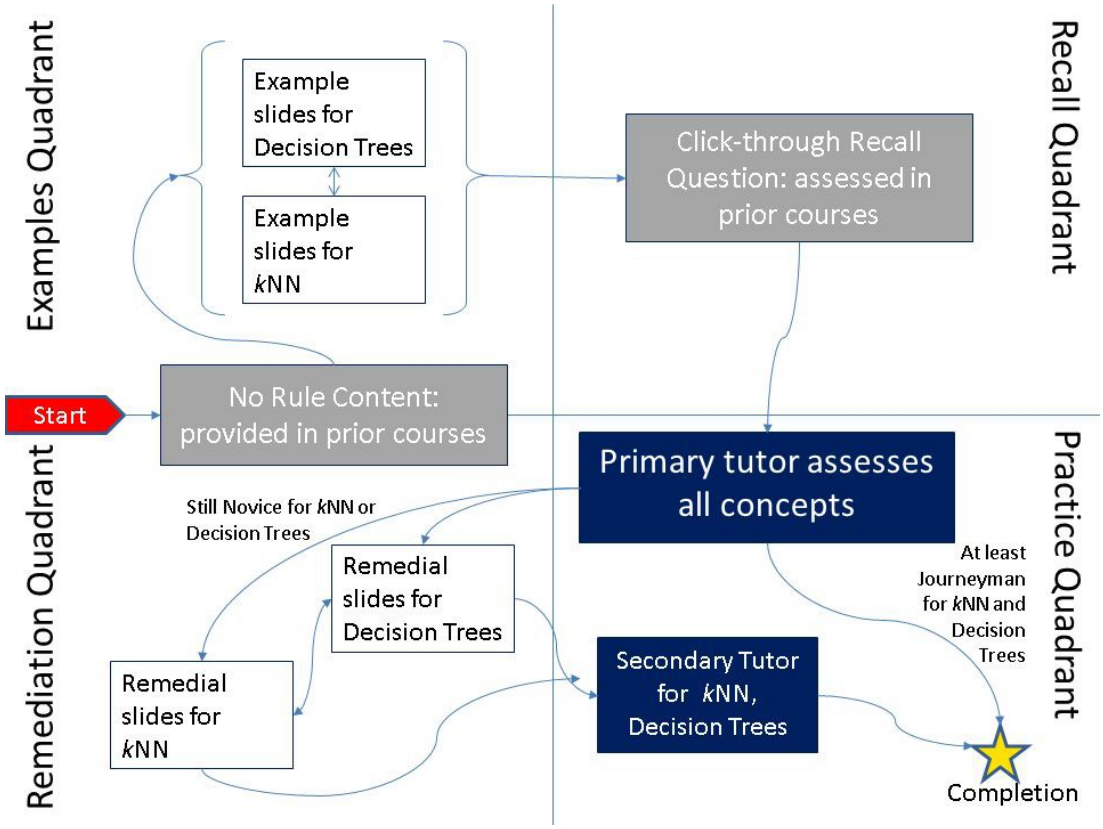


Figure 1: Week 1 Adaptive Courseflow object with CTAT tutors as Practice applications. The GIFT concepts and corresponding CTAT skills refer to the data mining algorithms Decision Trees and *k*NN (for *k*-Nearest Neighbor).

Even if this pattern of adaptive instruction - complex tutored problem-solving adaptively combined with remedial examples and remedial tutored problem solving - is simple, it extends what GIFT and CTAT easily do separately, and capitalizes on strengths of each tool. The complex tutored problem-solving activities authored with CTAT could not easily have been authored in GIFT, as GIFT does not support a non-programmer approach to creating user interfaces or adaptive inner loops, as CTAT does (Aleven et al., 2016). On the other hand, the adaptive interleaving of problem solving and declarative instruction, based on Merrill's quadrants, could not have been authored as easily in CTAT, because its standard adaptive outer loop option, namely, cognitive mastery based on Bayesian Knowledge Tracing (Corbett & Anderson, 1995), is geared towards problem solving only, without declarative instruction interleaved. CTAT does not represent the quadrant structure (an author would have to write a custom outer loop), and is not geared towards embedding external learning objects such as Powerpoint slides. We note that our proof-of-concept pattern of instruction realizes (in a new, more adaptive way) one of the Cognitive Tutor principles (Anderson, Corbett, Koedinger, & Pelletier, 1995; Koedinger & Corbett, 2006), namely, to "Provide instruction in the problem-solving context." In previous Cognitive Tutors, the declarative instruction was provided in the classroom (Koedinger, Anderson, Hadley, & Mark, 1997), or was embedded in the tutor as static text pages, though without adaptive sequencing.

STUDENT MODEL MAPPING

We saw two principal questions with respect to the semantics of GIFT's and CTAT's student models: How do key elements of CTAT's student model (mastery probabilities for KCs) correspond to the richer set of categories for representing knowledge in GIFT's student model? Second, how can CTAT's KC mastery probabilities be mapped

onto the three mastery levels used in GIFT's student model (Novice, Journeyman, Expert) for use with adaptive decisions, in a manner that respects the semantics of these categories, as intended by the GIFT designers?

First, a brief look at what these student models contain. GIFT (cf. the EMAP explanation at https://gifttutoring.org/projects/gift/wiki/Engine_For_Management_of_Adaptive_Pedagogy_(eMAP)_2018-1) decomposes expertise into concepts but also recognizes affective state and has a notion of behavior state. Concepts may be hierarchical, where a single overarching concept decomposes into a tree of finer-grained concepts, but this multi-level modeling is not required: concepts may instead be enumerated in a simple single-level list. Assessment of a student's mastery of each concept is maintained with respect to 1 or 2 measures, **Cognitive Knowledge** and **Cognitive Skill**. The latter is defined as the "ability to execute." For each concept GIFT maintains separate assessments of Cognitive Knowledge and Cognitive Skill as one of Novice, Journeyman or Expert. The assessment drives adaptive decisions within GIFT's Adaptive Courseflow object.

CTAT's student model is a set of independent knowledge components (KCs), also called skills. For each, CTAT records a probability that the student has mastered it, based on their prior performance; a single threshold (0.95 by default) indicates mastery. This technique for modeling students' knowledge, originally developed in Cognitive Tutors for personalized problem selection, tries to model especially procedural knowledge. Knowledge components are fine-grained: their scope can be refined empirically by observing error rates on questions thought to require the same knowledge (Aleven & Koedinger, 2013; Anderson et al., 1995).

For our proof-of-concept system, as an initial position we simply make a 1:1 correspondence between CTAT's knowledge components and the Cognitive Skill assessment of the lowest-level GIFT concepts (that is, the leaf concepts if the GIFT course uses hierarchical concept modeling). Both seem meant to capture procedural knowledge. To map CTAT probabilities onto GIFT's Novice-Journeyman-Expert levels, we let a GIFT author set probability ranges for the 3 levels in the GIFT Authoring Tool. We are still experimenting with the actual ranges to use for Novice, Journeyman and Expert, as it is hard to find a principled basis for this choice. We have considered equating the expertise level needed for promotion in the GIFT course with CTAT's mastery threshold, and we have set GIFT's Expert level provisionally to the 0.95 probability of mastery threshold in CTAT. But, so far, we have set the Journeyman level, again provisionally, at 0.75 probability. This initial approach may be simplistic, but it permits GIFT to use its full adaptive decision-making capabilities in Adaptive Courseflow objects that include CTAT LTI tools as practice applications. Our discussion below explores the limitations of our initial integration. Our recommendations suggest straightforward changes to GIFT that would permit finer-grained adaptive decisions.

TECHNICAL INTEGRATION

In this section, we describe how we implemented the one-way student model communication (from CTAT to GIFT), using the LTI interoperability standard. GIFT accommodates external learning activities via two different mechanisms. Heretofore, most integrations have required custom Java-language gateway programs that conform to an interface specified in the Domain Knowledge File (Domain Knowledge File). The use of Java on the client makes these programs inconvenient to deploy over the World Wide Web, however. Therefore, we decided to use GIFT's second mechanism for integrating external activities, namely, its implementation of the Learning Tools Interoperability (LTI) Tool Consumer interface. In prior work on our project (Aleven et al., 2018), ARL enabled GIFT to accommodate learning activities that adhere to the LTI v1.1.1 standard (IMS 2012). Figure 2 illustrates our use of this integration, where GIFT itself is an LTI Tool Provider to edX, due to yet earlier work on our project (Aleven et al. 2017).

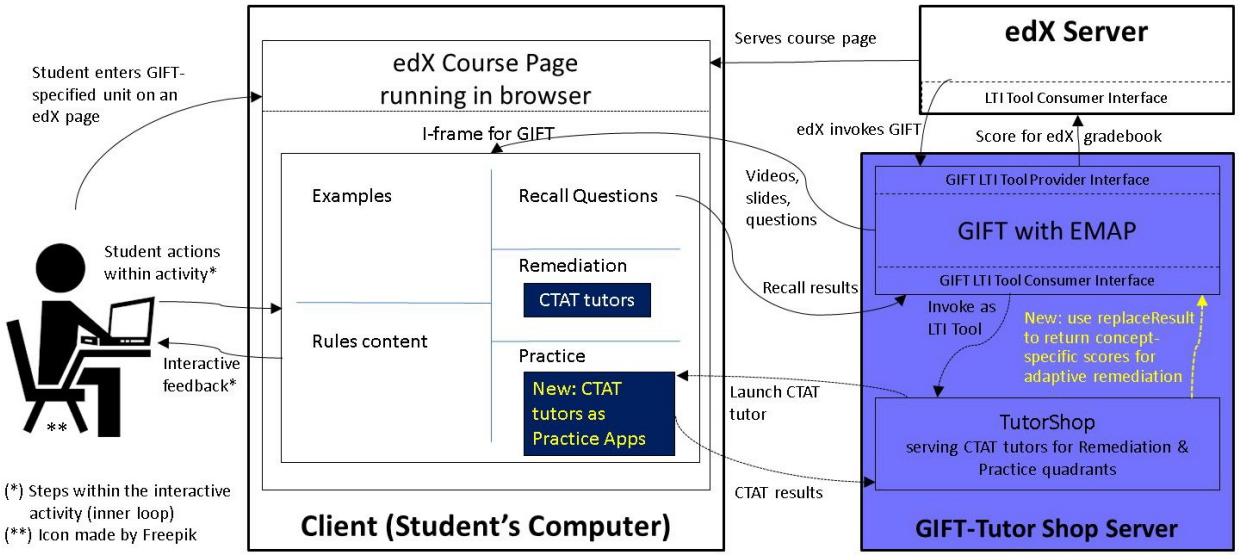


Figure 2: Architecture and control flow in the BDE MOOC week 1 course, with new features shown in yellow. GIFT is invoked by edX as an LTI Tool Provider and in turn launches CTAT tutors as Practice applications whose individual skill recalculations update concept-specific assessment levels in GIFT.

In our current work, CTAT tutors (introduced into the Practice quadrant) update GIFT's assessment of cognitive skill per concept. We make this update dependent on CTAT's calculation of probability of mastery of a corresponding knowledge component. To communicate the value, we use the LTI v1.1.1 specification's **replaceResult** request, by which a tool (here, CTAT), can return a single numeric score, with a label, to the tool consumer (GIFT). This single-score limitation presented a problem, since the GIFT course and the CTAT tutor generally track several concepts and knowledge components, respectively, each with different values. But we noted that the LTI specification permits the tool to issue the replaceResult request more than once, and we found that our off-the-shelf library implementations of the LTI interface permitted us to vary the labels on these requests. So, to permit CTAT to return multiple values to GIFT, we make special use of the current standard by allowing the tool to send many replaceResult requests, each with a concept-specific label and value.

In our integration, CTAT reports changes in skill mastery estimates while the student is working on a tutor, as soon as they happen, not just upon exiting the tutor. Part of the rationale is that the session could end abruptly at any time: the tutor might not have a chance to send final scores. A second reason is that the same student might be logged into GIFT in multiple sessions concurrently--even perhaps inadvertently. A common usage pattern on the World Wide Web is for a user to open a new browser tab to attend to some matter while in the midst of a task on another tab; after some time passes and more open tabs accumulate, it could be easy for a user to forget what tasks were in progress and so begin anew in a new tab on a site already active on another tab. Step-by-step reporting of student model updates helps student model values remain up-to-date for access from concurrent sessions.

After the time of our implementation (and the initial submission deadline for this paper), the LTI Assignment and Grade Services Specification v2.0 (https://www.imsglobal.org/spec/lti-ags/v2p0/) was published. This standard offers richer reporting capabilities. During our work it was neither finalized nor freely available; now we look forward to the development of support libraries to promote its adoption. Our v1.1.1 workaround was easy to implement with existing freely-available libraries, and it let us prototype a useful extension to GIFT's capabilities.

DISCUSSION

In this paper, we present a new GIFT/CTAT integration with one-way student model sharing between GIFT and CTAT, a novel feature, relative to our own prior work and to prior work in the field. Our approach is to exchange a

student model in one direction and to map student model values with 1:1 correspondence between the main student model elements. We present a proof-of-concept demonstration of this integration within one of the units of the summer 2019 edition of the edX course "Big Data and Education." In this unit, our integration supports new adaptive interleaving of complex problem solving and declarative instruction.

Our proof-of-concept represents a relatively simple special case of a more complex problem. First, it is limited in its curricular scope: it covers only a portion of one week's worth of instruction within the BDEMOOC, captured in one instance of GIFT's Adaptive Courseflow object. Nonetheless, similar extensions could be repeated in other course chapters.

Another limitation may be that in our project, we have pretty much a best case situation (albeit one that may not be uncommon) in which a single content development team works with all the different tools being integrated. This situation no doubt makes it easier to create student models whose elements map 1:1 than it would be otherwise. Things may be very different when integrating, post-hoc, two ITSs from different authors. Under those circumstances, the student models might not be as easy to align, and some form of ontological translation may be necessary. Perhaps even greater benefits from integration accrue when integrating existing systems; more content may be involved, for example. Then again, we do not know how likely that kind of integration scenario is.

The integration of the two student models may incur a certain level of what we might call semantic friction. Do thresholds on CTAT KC probabilities capture what the GIFT designers meant by the categories of Novice, Journeyman, and Expert? Some degree of semantic mismatch might be inevitable and perhaps unresolvable. We would like to think, however, that the current thresholds (set at .75 for Journeyman and .95 for Expert) align reasonably well with the intent of the GIFT designers, although some further scrutiny of this issue might be in order, informed perhaps by prior work in the Knowledge/Skills/Abilities (KSA) doctrine, on which the GIFT student model draws.

A further limitation is that we present no data to support the point that the current integration - although supported by instructional design principles - is actually benefiting students. The current work should be viewed as exploratory, focused mainly on technical issues. We are collecting data in this summer's run of the BDEMOOC, and plan to discover the different pathways that students take through the course materials. We do not expect however that the new adaptive pattern just in week 1 of the BDEMOOC will lead to measurably different outcomes (e.g., learning gains, retention rates). It may be better to wait with a more rigorous evaluation of student learning until more content has been moved into this (and similar) adaptive patterns.

A final limitation of the current integration is that it supports one-way communication of the student model only, namely, from CTAT to GIFT. Full integration would require two-way communication, so CTAT can be cognizant of information about the student inferred from performance in other GIFT activities, an interesting avenue for future work, as discussed below.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

In this paper, we solved technical issues regarding the one-way sharing and 1:1 mapping of student models, as one way in which one might integrate GIFT and CTAT. As a proof-of-concept, we demonstrated this integration by adaptive interleaving of problem-solving practice with remedial example studying and problem solving. These adaptive tutoring behaviors would be harder to author in either GIFT or CTAT alone. The main contribution of the work is that it demonstrates benefits of simple one-way student model sharing between ITS platforms, one of very few demonstrations in the literature of ITS integration. It demonstrates, as well, that integration of ITSs, generally viewed as both highly desirable and highly challenging, does not always need to be exceedingly difficult.

Our plans for future work are as follows: After running the edX course (planned for late spring and early summer 2019), we will analyze the course data to get a sense for the functioning of the new adaptive mechanisms in the

course that are made possible by the newly-implemented student model sharing. We will check how students' paths through the course changed and whether they became more effective, and to learn how well participation held up across adaptive transitions.

If the new v2.0 LTI standard with richer reporting capabilities gains broad adoption, we recommend its implementation in GIFT and in CTAT. Independent of that, we recommend two near-term enhancements and suggest some longer-term ideas. First, for the near term, it would be good to extend the GIFT Authoring Tool to permit concept-specific ranges for translating LTI numeric scores into GIFT's Novice-Journeyman-Expert expertise levels. That is, instead of a single slider to set a single tuple of Novice-Journeyman-Expert ranges for *all* concepts, it would be useful to be able to set these ranges *per* concept. Then authors would have flexibility to establish concept-specific performance expectations. Perhaps a way could be found that such flexibility would help reduce some of the semantic friction discussed above.

A second near-term recommendation for GIFT would be to enhance the Practice quadrant's content selection algorithm to account for cognitive skills already mastered. As mentioned above, when a student re-enters the Practice quadrant after remediation, GIFT currently considers only those applications associated with *all* concepts covered by the Practice phase. The student might avoid repetitive work if GIFT were able to choose a Practice application that covered only those concepts for which the student has not yet met expectations.

Further out, we note that GIFT's student model also includes Affective State. Recent work in CTAT has made it easy to integrate detectors that infer, from the student-tutor transaction stream, variables regarding student affect, unproductive persistence, disengaged behaviors such as gaming the system, and so forth (Holstein et al., 2018). It would be valuable, in the future, to investigate how these additional variables could be shared between CTAT and GIFT in a generalizable manner.

Another key direction would be to communicate the student model from GIFT back to CTAT, so the student model would be shared in bi-directional fashion, and the tutoring behavior in a CTAT tutor could adapt to what the student learned (or did not learn) in GIFT activities. Straightforward extensions of the LTI implementations of GIFT and CTAT (see readResult, below) would permit CTAT to receive the GIFT student model. If we assume, as in our current one-way integration, that GIFT skills would be mapped 1:1 to CTAT KCs, a key issue would be: How should we translate GIFT skill levels (novice, journeyman, expert) into CTAT KC probabilities? It seems inevitable that in the mapping we would lose "resolution." CTAT has greater precision in its student model values (which represent probabilities) than GIFT, a downside of GIFT's choice to distinguish only three mastery levels in its student model. A possible way around this loss of resolution might be for the two systems to each maintain their own student model, and to share student model *updates*, rather than the student model itself - so each system could update its student model based on events that happened in the other system. While that solution might maintain resolution, and perhaps avoid semantic friction of the kind described above, it would be more difficult to implement; we did not explore it.

In implementing two-way student model sharing, we may need to account for the possibility that a student may at times be working on two different activities simultaneously, as described above. If simultaneous activities share targeted knowledge components, then the updated student models sent from the Tool (e.g., CTAT) to the Consumer (GIFT) may clobber each other. This issue can be avoided by having the Tool first query the Consumer's student model for its current value(s) before computing an updated value and sending it back to the Consumer. Similarly, the Tool should query the Consumer's student model before and, based on the result of the query, update its own student model before using it for its own adaptive pedagogical decisions. CTAT would need some straightforward modifications to make these queries. In short, we see a clear path to two-way student model sharing between GIFT and CTAT, assuming that the 1:1 mapping of student model elements will remain appropriate.

Finally, the most exciting future work is exploring what new student experiences can be authored with the the new GIFT/CTAT integration. Some attractive scenarios might involve CTAT's cognitive mastery outer loop, which has been very well-studied in the EDM literature, and proven to be effective (Corbett, McLaughlin, & Scarpinatto, 2000), or more flexible interleaving of problem solving and declarative instruction. These scenarios might require additional

flexibility in, and authoring capabilities for, how GIFT adaptively traverses the quadrants in the Adaptive Courseflow object, and may benefit from ways of modeling links between procedural and conceptual knowledge, already represented in GIFT's student model, but not in CTAT's.

REFERENCES

Aleven, V., Baker, R., Blomberg, N., Andres, J.M., Sewall, J., Wang, Y., Popescu, O. (2017). Integrating MOOCs and Intelligent Tutoring Systems: edX, GIFT, and CTAT. In *Proceedings of the Fifth Annual GIFT Users Symposium*, May 2017.

Aleven, V., & Koedinger, K. R. (2013). Knowledge component approaches to learner modeling. In R. Sottilare, A. Graesser, X. Hu, & H. Holden (Eds.), *Design recommendations for adaptive intelligent tutoring systems* (Vol. I, Learner Modeling, pp. 165-182). Orlando, FL: US Army Research Laboratory.

Aleven, V., McLaren, B. M., Sewall, J., van Velsen, M., Popescu, O., Demi, S., Ringenberg, M., & Koedinger, K. R. (2016). Example-tracing tutors: Intelligent tutor development for non-programmers. *International Journal of Artificial Intelligence in Education, 26*(1), 224-269. doi:10.1007/s40593-015-0088-2.

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2017). Instruction based on adaptive learning technologies. In R. E. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 522-560). New York: Routledge.

Aleven, V., & Rosé, C. P. (2004). Towards easier creation of tutorial dialogue systems: Integration of authoring environments for tutoring and dialogue systems. In J. Mostow & P. Tedesco (Eds.), *Papers from the Workshop on Dialog-Based Intelligent Tutoring Systems*, held in conjunction with ITS 2004 (pp. 1-7). Maceió, Brazil.

Aleven, V., Sewall, J., Popescu, O., Sottilare, R., Long, R., & Baker, R. (2018). Towards adapting to learners at scale: Integrating MOOC and intelligent tutoring frameworks. In *Proceedings of the Fifth Conference on Learning @ Scale*, June 2018..

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. The *Journal of the Learning Sciences, 4*(2), 167-207.

Baker, R.S. (2016) Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence and Education, 26*(2), 600-614.

Baker, R.S. (under review) Some challenges for the next 18 years of learning analytics. Manuscript under review. Brusilovsky, P. (1995). Intelligent learning environments for programming: The case for integration and adaptation. In J. Greer (Ed.), Proceedings of the 7th World Conference on Artificial Intelligence in Education (AIED'95), (Vol. 95, pp. 1-8). Washington, DC: Association for the Advancement of Computing in Education.

Bull, S., & Kay, J. (2016). SMILI☺: A framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education, 26*(1), 293-331. doi:10.1007/s40593-015-0090-

Cai, Z., Graesser, A. C., & Hu, X. (2015). ASAT: AutoTutor script authoring tool. In R. Sottilare, A. Graesser, X. Hu, & K. Brawner (Eds.), *Design recommendations for adaptive intelligent tutoring systems* (Vol. III, Authoring Tools, Chap. 17, pp. 199-210). Orlando, FL: US Army Research Laboratory.

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

Corbett, A., McLaughlin, M., & Scarpinatto, K. C. (2000). Modeling student knowledge: Cognitive Tutors in high school and college. *User Modeling and User-Adapted Interaction, 10*, 81-108.

Holstein, K., Yu, Z., Sewall, J., Popescu, O., McLaren, B. M., & Aleven, V. (2018). Opening up an Intelligent Tutoring System development environment for extensible student modeling. In C. P. Rosé, R.

Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Proceedings, 19th International Conference on Artificial Intelligence in Education, AIED 2018* (Part 1, pp. 169–183). Cham, Switzerland: Springer. doi: 10.1007/978-3-319-93843-1_13

IMS Global Learning Tools Interoperability™ Implementation Guide Final Version 1.1.1 (2012), at https://www.imsglobal.org/specs/ltiv1p1p1/implementation-guide.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*(1), 30-43.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61-78). New York: Cambridge University Press.

Koedinger, K. R., Suthers, D. D., & Forbus, K. D. (1998). Component-based construction of a science learning space. In B. P. Goettl, H. M. Halff, C. L. Redfield, & V. J. Shute (Eds.), Intelligent Tutoring Systems, Fourth International Conference, ITS '98 (pp. 166-175). Lecture Notes in Computer Science 1452. Berlin: Springer Verlag.

Sottilare, R., Long, R. and Goldberg, B. (2017). Enhancing the Experience Application Program Interface (xAPI) to improve domain competency modeling for adaptive instruction. *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, April 2017, 265-268.

Goldberg, B., Hoffman, M. and Tarr, R. (2015). Authoring Instructional Management Logic in GIFT Using the Engine for Management of Adaptive Pedagogy (EMAP). *Design Recommendations for Intelligent Tutoring Systems: Authoring Tools and Expert Modeling Techniques*, 319-334.

Mitrovic, A., Martin, B., Suraweera, P., Zakharov, K., Milik, N., Holland, J., & McGuigan, N. (2009). ASPIRE: an authoring system and deployment environment for constraint-based tutors. *International Journal of Artificial Intelligence in Education, 19*(2), 155-188.

Razzaq, L., Patvarczki, J., Almeida, S. F., Vartak, M., Feng, M., Heffernan, N. T., & Koedinger, K. R. (2009). The Assistment Builder: Supporting the life cycle of tutoring system content creation. *IEEE Transactions on Learning Technologies, 2*(2), 157-166.

Sottilare, R. and Gilbert, S. (2011). Considerations for tutoring, cognitive modeling, authoring and interaction design in serious games. Authoring Simulation and Game-based Intelligent Tutoring workshop at the Artificial Intelligence in Education Conference (AIED) 2011, Auckland, New Zealand, June 2011.

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

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

Sottilare, R., Holden, H., Goldberg, B., & Brawner, K. (2013). The Generalized Intelligent Framework for Tutoring (GIFT). In Best, C., Galanis, G., Kerry, J. and Sottilare, R. (Eds.) *Fundamental Issues in Defence Simulation & Training*. Ashgate Publishing.

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

VanLehn, K. (2016). Regulative loops, step loops and task loops. *International Journal of Artificial Intelligence in Education, 26*(1), 107-112. doi:10.1007/s40593-015-0056-x

Vygotsky, L.S. (1978). Mind in Society: The development of higher psychology processes. Cambridge MA: Harvard University press.

ABOUT THE AUTHORS

Dr. Vincent Aleven *is a Professor of Human-Computer Interaction at Carnegie Mellon University. His research focuses on adaptive learning technologies, with a grounding in cognitive theory and theories of self-regulated learning. He and his colleagues have created the Cognitive Tutor Authoring Tools (CTAT), which has been used to build a very wide range of intelligent tutoring systems, many of which have been used in real educational settings.*

Mr. Miggy Andres-Bray *is a 4th year PhD student at the University of Pennsylvania, working on the MOOC Replication Framework.*

Dr. Ryan S. Baker *is Associate Professor at the University of Pennsylvania's Graduate School of Education and directs the Penn Center for Learning Analytics. He teaches the MOOC Big Data and Education.*

Mr. Rodney Long *is a Science and Technology Manager at the Army's Soldier Center - Simulation and Training Technology Center in Orlando, Florida and conducts research in adaptive training technologies.*

Dr. Octav Popescu *is a Senior Research Programmer/Analyst in Carnegie Mellon's Human-Computer Interaction Institute, where he is in charge of TutorShop, the Learning Management System part of the Cognitive Tutor Authoring Tools project.*

Mr. Jonathan Sewall *is a project director at CMU's Human-Computer Interaction Institute, working on the Cognitive Tutor Authoring Tools (CTAT).*

Dr. Robert Sottilare *leads adaptive training research within ARL's Learning in Intelligent Tutoring Environments (LITE) Lab and is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT).*

THEME IV:
INSTRUCTIONAL MANAGEMENT AND TRAINING EFFECTIVENESS

Towards Data-Driven Tutorial Planning for Counterinsurgency Training in GIFT: Preliminary Findings and Lessons Learned

Randall Spain[1], Jonathan Rowe[1], Benjamin Goldberg[3] , Robert Pokorny[2], Bradford Mott[1] and James Lester[1]
North Carolina State University[1], Intelligent Automation, Inc.[2], U.S. Army Combat Capabilities Development Command Soldier Center - STTC[3]

INTRODUCTION

Adaptive instructional systems (AISs) guide student learning experiences by tailoring instruction based on the individual goals, needs, and preferences of learners in the context of domain learning objectives (Sottilare, Barr, Robson, Hu & Graesser, 2018). A critical feature of AISs is the capability to dynamically guide and scaffold student learning. Leveraging recent advances in artificial intelligence and machine learning, it is possible to tailor training and educational experiences to individuals and teams of learners. Tutorial planning is a critical component of AISs, controlling how scaffolding is structured and delivered to learners. Tutorial planners operate at multiple levels, selecting problems for learners to solve and delivering tailored hints and feedback about specific problems. While research shows AISs can help to improve learning gains in many domains, devising computational models that determine when to scaffold, what type of scaffolding to deliver, and how scaffolding should be realized, is a critical challenge for the field.

Over the past several years the Generalized Intelligent Framework for Tutoring (GIFT) has emerged as a key exemplar of how these challenges in developing ITSs can be addressed at scale (Sottilare, Brawner, Goldberg, & Holden, 2012; Sottilare, Brawner, Sinatra, & Johnston, 2017). GIFT is an open-source domain-independent software framework for designing, deploying, and evaluating adaptive training systems. GIFT provides instructors with a suite of web-based tools for rapidly creating intelligent tutors, and it is linked to several ongoing research efforts to devise methods for automating key elements of the adaptive training authoring process. Many of these tools are available through GIFT's Course Creator, which provides a drag-and-drop interface for devising adaptive training experiences across a range of domains.

In this paper, we describe results from a research program that aims to devise data-driven tutorial planning policies that can be used in GIFT to present learners with adaptive remediation. In particular, we present preliminary results from a study involving over 500 learners who completed an approximately two-hour hypermedia training course that taught doctrinal concepts associated with counterinsurgency (COIN) and stability operations. The course leverages several unique enhancements to GIFT's Engine for Management of Adaptive Pedagogy (EMAP) including a newly developed remediation module that presents learners with passive, active, or constructive forms of remedial feedback. The remediation activities are based on the ICAP framework for active learning (Chi, 2009) which predicts that interactive remediation (e.g., peer dialogue) is more effective for learning than constructive remediation (e.g., writing an explanation), constructive remediation is more effective than active remediation (e.g., reading and highlighting a passage), and active remediation is more effective than passive remediation (e.g., reading a passage without doing anything else). Our analyses address several fundamental questions that are essential for developing effective reinforcement learning policies. We also describe lessons learned from deploying the training course on Amazon's Mechanical Turk (MTurk) and utilizing GIFT's Event Reporting Tool to extract data in support of reinforcement learning analysis. The paper concludes with a discussion of upcoming plans to devise tutorial policies using reinforcement learning techniques, as well as future directions for incorporating these policies in GIFT to enhance its ability to provide learners with effective and efficient adaptive remediation in future courses.

RESEARCH CONTEXT

A significant challenge that authors face when designing AISs is determining when to scaffold learners, what type of scaffolding to deliver, and how scaffolding should be realized. One reason for this challenge is the wide range of pedagogical strategies and tactics that can be implemented in AISs, as well as a lack of empirically grounded guidance about the relative contribution of different adaptive interventions on learning outcomes (Durlach & Ray, 2011). Another challenge facing adaptive course designers is that rules that drive adaptive pedagogical decisions often must be manually engineered, which can significant- ly increase the time required to author adaptive instructional materials (Aleven, McLaren, Sewall, & Koedinger, 2009; Sottilare, 2015).

Recent developments in artificial intelligence and machine learning have introduced opportunities to reduce the authoring burden of AISs by devising data-driven tutorial planning policies that can automatically control how pedagogical support is structured and delivered to learners to create personalized learning experiences (Rowe & Lester, 2015; Williams et al., 2016; Zhou, Wang, Lynch, & Chi, 2017). Tutorial planning is an important component of ITSs that controls how instructional interventions are structured and delivered at the macro-level (e.g., selecting problems for learners to solve) and micro-level (e.g., delivering tailored hints and feedback about specific problems). Tutorial planning techniques are complementary to advances in intelligent tutoring system authoring, including authoring tools implemented in GIFT, to address the challenges inherent in constructing adaptive training materials.

Reinforcement learning techniques have shown promise for automatically inducing tutorial planning rules that optimize student learning outcomes and do not require pedagogical rules to be manually programmed or demonstrated by expert tutors. Reinforcement learning is a category of machine learning that centers on devising software agents that perform actions in a stochastic environment to optimize some concept of numerical reward (Sutton & Barto, 1998). In reinforcement learning, the agent induces a control policy by iteratively performing actions and observing their effects on the environment and accumulated rewards. Tutorial planning can be formalized as a reinforcement learning task in which the tutor (i.e., agent) aims to make pedagogical decisions (i.e., actions) that will affect its environment (i.e., the trainee and his/her learning environment) to optimize student learning outcomes (i.e., rewards). In our case, the pedagogical decisions are choosing between ICAP-inspired remediation activities, and the tutorial planner's objective is to optimize student learning in an adaptive hypermedia-based training course for COIN.

Because reinforcement learning techniques are data-intensive, a critical goal of our study was to obtain a large dataset consisting of trainee responses to different types of instructional remediation activities. To meet this objective, we developed an adaptive hypermedia-based training course in GIFT that builds upon materials from the UrbanSim Primer. Originally developed by the USC Institute for Creative Technolo-gies, the UrbanSim Primer is a hypermedia-based learning environment that provides direct instruction on key concepts and principles of COIN doctrine. Our GIFT-based version of the UrbanSim Primer course was designed to: (1) contain numerous opportunities for learners to receive instructional remediation; (2) be deployable through online crowdsourcing platforms, which enabled efficient distribution to many learners for data collection purposes; (3) enact an exploratory (i.e., random) remediation policy in order to broadly sample the space of possible pedagogical decisions; (4) assess learning gains using pre-and post- knowledge tests, and (5) collect trace data from participants as they interacted with the training course (i.e., how many times learners received remediation, how long they spent interacting with the different forms of remediation, correctness of responses, helpfulness ratings of remedial content, etc.) which would enable exploration of different state representations and reward functions for inducing reinforcement learning-based tutorial policies.

In the following sections, we describe the results of a large human subject's study that we recently completed as well as the preliminary analyses that serve as prerequisites for developing tutorial policies using reinforcement learning techniques. The research questions guiding our initial set of analyses included: How effective was the course in promoting learning gains? How frequently did learners receive remediation? How long did learners spend interacting with each form of remediation? Which form of remediation was most effective for helping learners overcome an impasse?

METHODOLOGY

Participants

To meet our goal of facilitating broad distribution to many learners, we recruited participants through Amazon's MTurk platform. Participants were required to be at least 18 years of age, reside in the United States, have completed at least 50 MTurk tasks, and have obtained a task completion success rate of at least 95% to be eligible for the study. A total of 533 participants (42% female, ages ranged from 18 - 65) completed the training course, which lasted approximately 2 hours. Participants received $8 for completing the full training course. Thirty-five percent of participants had a bachelor's degree, 25% had some college education, 11% had a master's degree, and 11% had a high school diploma. Two percent of the sample reported being extremely familiar with COIN principles and doctrine; 12% reported being extremely interested in learning about COIN topics.

Hypermedia Training Course

The hypermedia-based training course was based upon materials from the UrbanSim Primer. The course was authored in GIFT and organized into 4 chapters. Each chapter contained a series of short videos, recall questions, and remedial training content designed to teach common themes, terminology and principles of COIN operations (Rowe, Spain, Pokorny, Mott, Goldberg, & Lester, 2018). The videos were approximately 90 seconds and covered topics such as "Identifying the center of gravity in COIN operations", "Defining intelligence preparation for the battlefield", and "Understanding lines of effort in COIN operations." The recall questions, which were presented in multiple choice format, assessed the content covered in the videos. The remediation interventions were structured according to Chi's ICAP framework (Chi, 2009) and required students to either passively, actively, and constructively engage with remedial feedback upon missing a quiz question. The hypermedia course also included a set of web-based surveys designed to collect information about participants' age, education, interest in counterinsurgency operations and military science topics, and goal orientation, as well as parallel forms of a 12-item pre-and posttest that measured knowledge of COIN topics, terminology, and principles. The hypermedia course contained a total of 12 multimedia videos, 39 multiple-choice recall questions, and 168 ICAP inspired remediation files. Participants advanced through the training course at their own pace and were not allowed to review previously completed lessons or videos.

Procedure

A brief description of the study was posted on the MTurk website. Participants who were interested in the study reviewed and electronically signed an informed consent form that described the study's purpose, risks, benefits, and compensation requirements. Afterward participants proceeded to the training course which was hosted on the cloud-based instance of GIFT.

The course began with a general message that welcomed participants to the training course. Following this introduction, participants completed a demographic questionnaire that gathered information about their age, years of education, and familiarity with COIN topics and concepts. Then, they completed a goal orientation questionnaire that measured task-based and intrinsic motivation to learn (Elliot & Murayama, 2008) followed by a 12-item pretest that measured prior knowledge of COIN principles and terminology. After completing the pre-training surveys, participants began the adaptive hypermedia COIN training course. Participants watched a series of narrated videos that covered lesson topics such as the importance of population support, processes for intelligence gathering, and issues in successful COIN operations. After each video, participants completed a series of recall questions that consisted of single or multi- concept review items that aligned with the course's learning objectives. Single concept review questions required learners to recall and apply concepts presented within the video lesson. Multi-concept review questions required learners to demonstrate a deeper understanding of course material by integrating concepts from multiple lessons. Following a missed question, participants received ICAP-inspired remediation that required

them to either: (1) *passively* re-read the narrated content that was just presented in the lesson video; (2) re-read the video content and *actively* highlight the portion of text that answered the recall question that was just missed; or (3) re-read the text and *constructively* summarize the answer to the recall question in their own words. The active and constructive remediation prompts also included expert highlighting/summaries that asked students to self-evaluate the accuracy of their responses (see Figure 1).

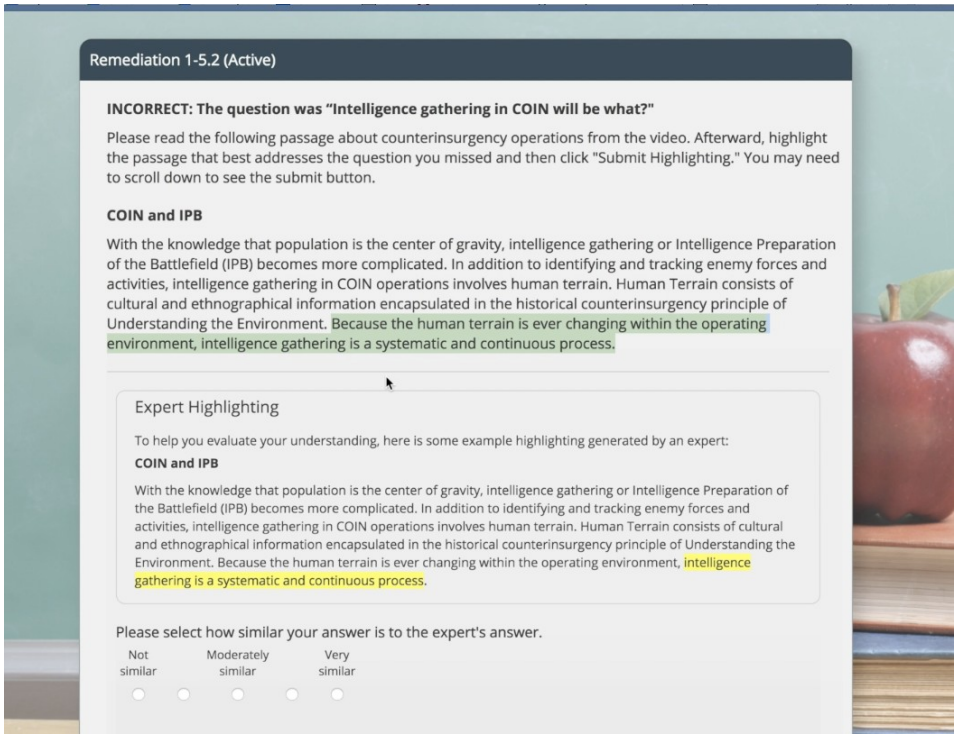


Figure 1. Example active remediation activity.

The course also included a "no remediation" prompt that only provided students with minimal feedback before being asked to re-answer the quiz question. The course used a random assignment policy that determined whether students received passive, active, constructive, or no remediation after each incorrect item response. Students continued to receive remediation until they demonstrated concept mastery (i.e., correctly answering the recall question). In addition to the ICAP-inspired remediation prompts, the training course also monitored how long students engaged with the video-based lessons and provided prompts to those participants who advanced through the videos too quickly or too slowly.

Upon finishing the final lesson and quiz, participants completed a series of post-training surveys that included a multiple-choice posttest to measure retention of the concepts and principles presented in the training and a short questionnaire to collect opinions about the training experience. After completing these activities, participants received a debriefing message, they were thanked for their participation, and they received a unique completion code that could be used to verify course completion through the MTurk website.

PRELIMINARY RESULTS

A goal of the overall research program is to investigate the benefits of different tutorial interventions for improving student learning in adaptive training environments. Towards this goal, and prior to developing any reinforcement learning-based policies, we first conducted a set of preliminary analyses to identify how well participants performed

in the course, how often learners received remediation, and how long, on average, learners spent interacting with the different intervention forms.

Learning Gains

Participants' pre-and posttest scores as well as normalized learning gains were analyzed in order to determine if the course was effective in promoting participants' knowledge of COIN concepts, terminology, and principles. Scores from the 12-item pretest revealed that participants had low prior knowledge of the concepts covered in the course (M= 4.29, SD = 2.20). A post hoc analysis using a two-sample test t- test indicated that post test scores were significantly higher (M= 8.22, SD = 3.11) than pre-test scores, $t(509) = 30.79$, $p <$. 001. An analysis of participants' normalized learning gains showed the course was effective in meeting its instructional objectives (M = .52; SD = .11) and that participants benefited from completing the course.

Remediation

Next, we examined how often participants received remediation in the course. Reinforcement learning techniques are data intensive, and therefore it is critical that the dataset contain a large number of remedi- ation instances to broadly sample the space of possible tutorial interventions and support inducing data- driven tutorial policies. Results showed that the training corpus included a total of 5,189 instances of remediation. Individual participants typically received multiple instances of remediation in the range of 1 to 113 (M = 10.08, SD = 12.58). Although the course was designed to implement a randomized control policy, frequency statistics showed that 40% of all remediation interventions were active-interventions, 40% were constructive, 10% were passive, and 10% were no-remediation. A closer inspection of the remediation data showed that 5% of the sample received only one instance of remediation (i.e., partici- pants missed only one recall question and therefore received only one remediation intervention) and that 75% of the sample received up to 10 instances of remediation.

Following these analyses, we analyzed participants' completion times for each form of remediation to determine whether participants spent more time completing the constructive and active remediation activities, which were designed to evoke more cognitive engagement, compared to the passive remediation activities, which were designed to be less engaging. Our analysis showed that participants spent the most time completing the constructive remediation activities (M = 75.96; SD = 32.30), a moderate amount of time completing the active remediation activities (M = 44.26; SD = 15.25), and the least amount of time viewing the passive remediation content (M = 27.64; SD = 21.60).

Further analyses showed participants spent less time on the constructive ($r(39) = -.63$, $p <$. 001) and active remediation ($r(38) = -.58$, $p < .001$) activities as they progressed through the training course, but there were no significant decreases in viewing time across passive remediation interventions ($r(39) = -.20$, $p = .23$). These results suggest that participants spent increasingly less time completing the constructive and active remediation activities as they progressed through the training course (Figure 2). Notably, participants spent almost 2 minutes, on average, completing constructive remediation activities when the training course began. By the end of the third chapter participants spent roughly a minute completing the constructive remediation activities, and by the conclusion of the course they were spending approximately 40 seconds completing these activities. A similar, albeit less pronounced trend is evident for the active remediation activities as well. These data suggest participants may have grown fatigued with the more cognitively engaging forms of remediation as the course progressed.

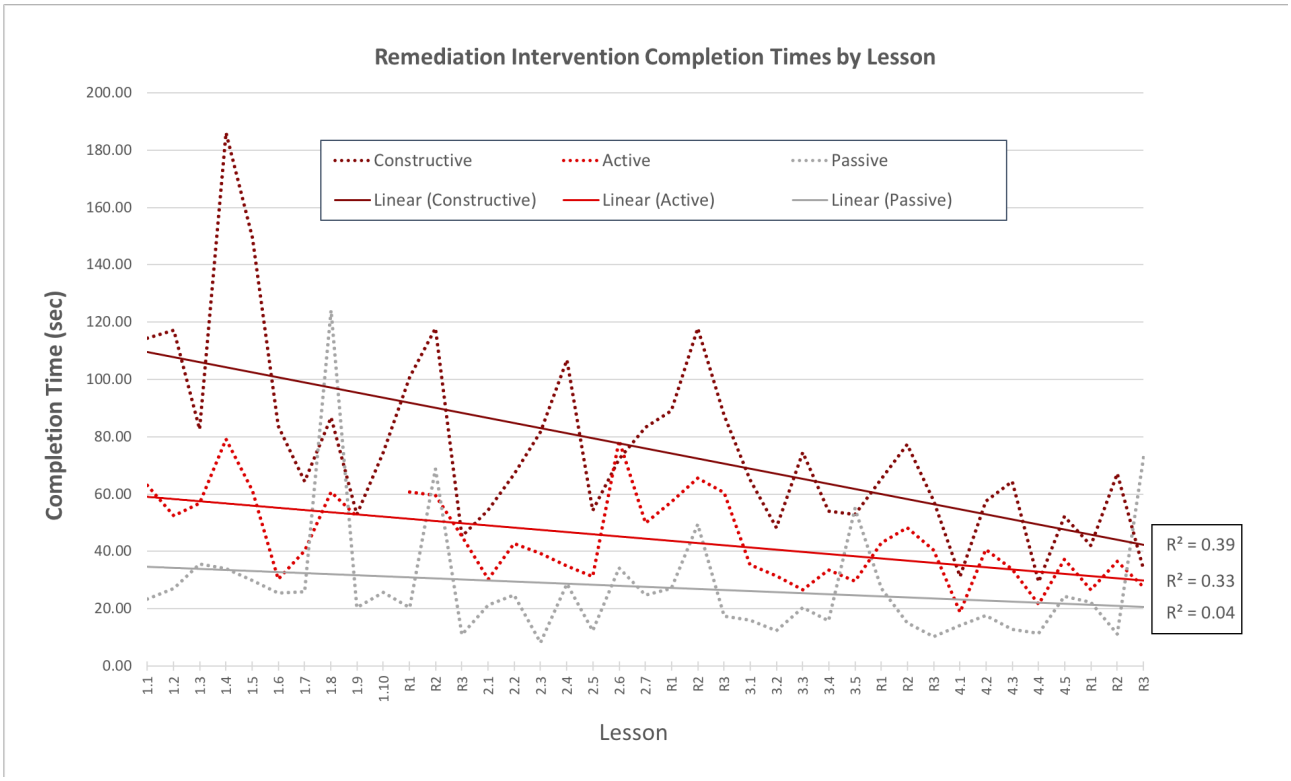
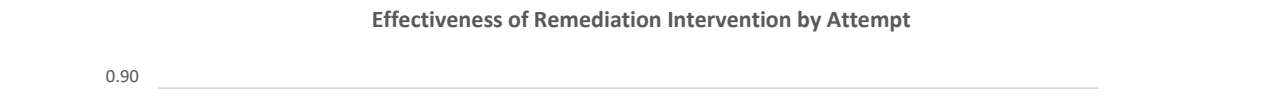


Figure 2: Remediation completion time across course lessons.

Finally, we conducted a set of exploratory analyses to identify which form of remediation was most effective at helping participants overcome an impasse for a missed recall item. We operationally defined *remediation effectiveness* as the proportion of cases in which participants correctly answered a recall question after receiving a given type of remediation (constructive, active, passive, none). We calculated remediation effectiveness in terms of the first, second, and third remediation instances delivered following missed attempts on a given recall question. As previously noted, participants continued to receive remediation until they demonstrated concept mastery. So, if a student missed a recall question, they continued to receive remediation until they answered the question correctly. By examining remediation effectiveness over successive attempts we aimed to identify trade-offs in remediation effectiveness that may have occurred as participants transitioned from one unsuccessful remediation attempt to another. The ICAP model predicts that constructive remediation should be more effective than active remediation at helping students overcome an impasse, and active remediation should be more effective than passive remediation. However, there could be tradeoffs between these different forms, as evident in the previous set of analyses that showed participants spent less time completing constructive and active remediation activities as they progressed through the course.

Our results generally supported the predictions of the ICAP model. Constructive remediation appeared to be more effective compared to active remediation at helping students overcome an impasse after one round of remediation, active remediation appeared to be more effective than passive, and passive remediation appeared to be more effective than no remediation (Figure 3). For cases in which participants received two rounds of remediation before correctly answered a recall question, constructive and active remediation appear to be the most effective form of remediation. Interestingly, presenting no remediation appeared to be more effective than presenting passive remediation. For cases in which participants correctly answered a recall question after the third remediation attempt, active remediation appeared to be more effective, followed by constructive remediation. There did not appear to be a major difference between passive and active remediation in terms of effectiveness.



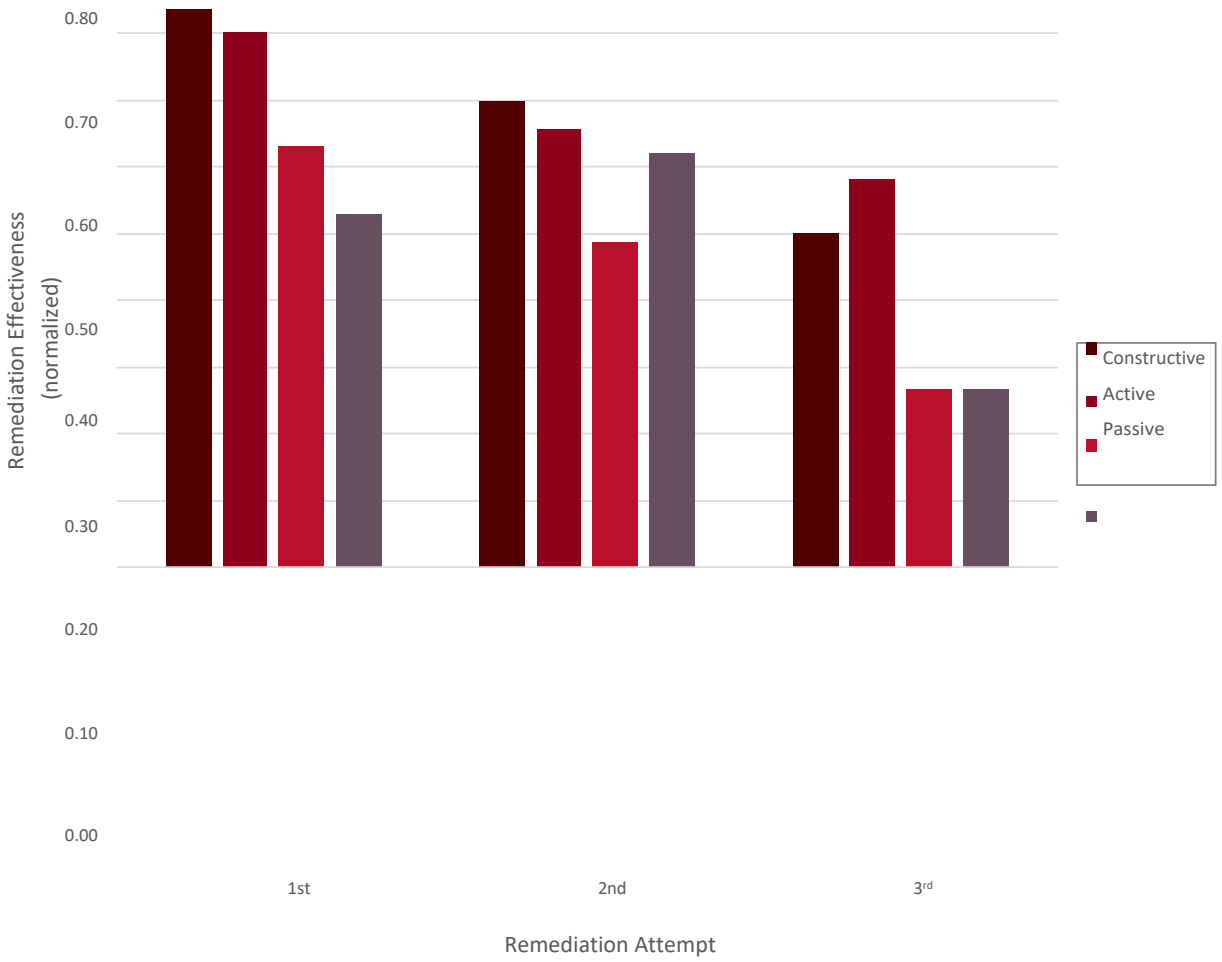


Figure 3: Remediation effectiveness across remediation attempts.

DISCUSSION AND LESSONS LEARNED

Our initial results are promising, and they suggest that the training corpus we collected using GIFT and Mechanical Turk contains a sufficient number of remediation interventions to explore data-driven tutorial planning models with reinforcement learning techniques. In addition, participants' interaction behaviors with the remediation appeared to align with expected outcomes of the ICAP framework. Notably, participants interacted with the constructive and active remediation content longer than the passive remediation content. The constructive and active remediation content also appeared to be more effective in helping learners overcome impasses during the course. Importantly, participants' knowledge of COIN concepts improved from pretest to post-test.

To our knowledge this is the first time GIFT has been used with an online crowdsourcing platform to collect a large corpus of training data for machine learning analysis. By using MTurk, we were able to collect data from over 500 users over the course of a 4-week timespan. As we conducted the study, we adopted several best practices to ensure data were collected in an efficient and effective manner. First, we used multiple batches to collect data (approximately 15) and limited batch sizes to approximately 50 slots. This served two primary purposes: (1) it made monitoring the course and completion rates more manage- able for the research team, and (2) it allowed us to make changes to the course based on user feedback. Second, we closely monitored the email account associated with the MTurk profile and responded to all inquiries regarding the course. The MTurk user community is extremely

responsive and forthcoming with feedback. Many participants shared recommendations for course improvements during our pilot testing phases as well as during testing. Participants frequently used the account's email address to notify us if they experienced trouble completing the course or if they ran into difficulties submitting the completion code. By providing timely responses to participants, we were able to quickly address any unforeseen issues that participants experienced and maintain a high rating on many of the MTurk community forums where users review and rate MTurk tasks.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

Recent advances in ITS authoring tools, as well as data-driven tutorial planning, are showing significant progress toward reducing the effort required to create personalized learning experiences. A key next step is the development of computational methods and tools for automatically inducing pedagogical models that dynamically tailor learning experiences across different domains and learning environments. In this paper, we have reported preliminary results from a study conducted using GIFT and the Mechanical Turk crowdsourcing platform that was designed to collect a training corpus for inducing tutorial planning models in a hypermedia-based course using reinforcement learning techniques. Results suggest that ICAP-inspired feedback and remediation in the GIFT-based course broadly follows trends predicted by the ICAP model concerning instructional design and student cognitive engagement. However, results also suggest that the effectiveness of ICAP-inspired remediation may change over time and under different conditions, pointing toward the need for data-driven tutorial policies to control how and when different forms of remediation are delivered to learners. These findings set the stage for investigating the application of reinforcement learning techniques to automatically induce tutorial policies for controlling how and when ICAP-inspired remediation is delivered to learners.

There are several promising directions for future research and development of GIFT. One recommendation is to expand the capability of Event Reporting Tool (ERT), which provides researchers with a means for extracting key data from users' interaction logs. The ERT produces a record of all events that occurred during the GIFT session. Some of these events specify what the learner did; other events result from GIFT processing. While the log file from a GIFT session captures the interactions of GIFT, the log file must be transformed into another file in order to make it useful for analysis of learning effectiveness. Our team is currently working on an open source tool that will allow researchers to transform data from the ERT into a format that is amenable to reinforcement learning analysis. As GIFT's user base continues to expand, it will become critically important to ensure researchers can easily access and analyze log data to investigate the effectiveness of different instructional inventions and tutorial policies.

ACKNOWLEDGEMENTS

The authors wish to thank Mike Hoffman for providing extensive assistance in developing GIFT software features that enabled this research. The research described herein has been sponsored by the U.S. Army Research Laboratory under cooperative agreement W911NF-15-2-0030. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

Aleven, V., McLaren, B., Sewall, J., & Koedinger, K. R. (2009). Example-tracing tutors: A new paradigm for intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, *19*(2), 105.

Chi, M.T.H. (2009). Active-Constructive-Interactive: A conceptual framework of differentiating learning activities. *Topics in Cognitive Science*, 1(1), 73-105.

Durlach, P. J., & Ray, J. M. (2011). *Designing adaptive instructional environments: Insights from empirical evidence*. Technical Report (1297) U.S. Army Research Institute for the Behavioral and Social Sciences; Arlington, VA.

Rowe, J., & Lester, J. (2015). Improving Student Problem Solving in Narrative Centered Learning Environments: A Modular Reinforcement Learning Framework. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence in Education,* pp. 419-428.

Rowe, J., Spain, R., Pokorny, B., Mott, B., Goldberg, B., & Lester, J. (2018). Design and development of an adaptive hypermedia course for counterinsurgency training in GIFT: Opportunities and lessons learned. *Proceedings of the Sixth Annual GIFT User Symposium* (GIFTSym6), pp. 229-239.

Sottilare, R. (2015). Challenges to Enhancing Authoring Tools and Methods for Intelligent Tutoring Systems. In R. Sottilare, A. Graesser, X. Hu, and K. Brawner (Eds.). (2015). Design Recommendations for Intelligent Tu- toring Systems: Volume 3 - Authoring Tools and Expert Modeling Techniques. Orlando, FL: U.S. Army Research Laboratory.

Sottilare, R., Barr, A., Robson, R., Hu, X., & Graesser, A. (2018). Exploring the opportunities and benefits of standards for adaptive instructional systems (AISs). In Proceedings of the Adaptive Instructional Systems Workshop in the Industry Track of the 14th International Intelligent Tutoring Systems (ITS) Conference, pp. 49-53.

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). The generalized intelligent framework for tutoring (GIFT). *U.S. Army Research Laboratory–Human Research & Engineering Directorate (ARL- HRED)*.

Sottilare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). An updated concept for a generalized intelligent framework for tutoring (GIFT). Retrieved from www.gifttutoring.org

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: an introduction MIT Press. *Cambridge, MA*. Williams, J. J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K., Lasecki, W., & Heffernan, N. (2016). AXIS: Generating Explanations at Scale with Learner sourcing and Machine Learning. Proceedings of the 3rd *Annual ACM Conference on Learning at Scale*, Edinburgh, SCT, pp. 379-388.

Zhou, G., Wang, J., Lynch, C., & Chi, M. (2017). Towards closing the loop: Bridging machine-induced pedagogical policies to learning theories. In *Proceedings of the Tenth International Conference on Educational Data Mining* (pp. 112-119).

ABOUT THE AUTHORS

Dr. Randall Spain *is a Research Psychologist in the Center for Educational Informatics at North Carolina State University where he uses principles, theories, and methods of applied psychology (human factors, educational psychology, personnel psychology, experimental psychology, and psychometrics) to design and evaluate the impact of advanced training technologies on learning and performance. He has conducted training and human factors research for the Department of Defense and the Department of Homeland Security for the past 10 years with a focus on adaptive training, performance assessment and measurement, user modeling and human-automation interaction. Dr. Spain is a PhD graduate from Old Dominion University's Human Factors Psychology program and serves on the editorial board for Military Psychology.*

Dr. Jonathan Rowe *is a Research Scientist in the Center for Educational Informatics at North Carolina State University, as well as an Adjunct Assistant Professor in the Department of Computer Science. He received the PhD and MS degrees in Computer Science from North Carolina State University, and a BS degree in Computer Science from Lafayette College. His research is in the areas of artificial intelligence and human-computer interaction for advanced learning technologies, with an emphasis on game-based learning environments, intelligent tutoring systems, multimodal learning analytics, user modeling, educational data mining, and computational models of interactive narrative generation. Dr. Rowe also serves on the editorial boards of the International Journal of Artificial Intelligence in Education and IEEE Transactions on Learning Technologies.*

Dr. Benjamin Goldberg *is a senior researcher in the Learning in Intelligent Tutoring Environments (LITE) Lab at the Combat Capabilities Development Command (CCDC) Soldier Center, Simulation and Training Technology Center (STTC) in Orlando, FL. He has been conducting research in modeling and simulation for the past eight years with a focus on adaptive learning and how to leverage artificial intelligence tools and methods for adaptive computer-based instruction. Currently, he is the LITE Lab's lead scientist on instructional strategy research within adaptive training environments. Dr. Goldberg holds a PhD from the University of Central Florida in Modeling & Simulation.*

Dr. Robert Pokorny *is Principal of the Education and Training Technologies Division at Intelligent Automation, Inc. He earned his PhD in Experimental Psychology at the University of Oregon in 1985, and completed a postdoctoral appointment at the University of Texas at Austin in Artificial Intelligence. His first position after completing graduate school was at the Air Force*

Research Laboratory, where he developed methodologies to efficiently create intelligent tutoring systems for a wide variety of Air Force jobs. At Intelligent Automation, Bob has led many cognitive science projects, including adaptive visualization training for equipment maintainers, and an expert system approach for scoring trainee performance in complex simulations.

Dr. James Lester *is Distinguished Professor of Computer Science at North Carolina State University, where he is Director of the Center for Educational Informatics. His research centers on transforming education with technology-rich learning environments. With a focus on adaptive learning technologies, his research spans intelligent tutoring systems, game-based learning environments, affective computing, and tutorial dialogue. The adaptive learning environments he and his colleagues develop have been used by thousands of students in K-12 classrooms. He received his PhD in Computer Science from the University of Texas at Austin in 1994. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI).*

Towards Accelerated Learning Pedagogical Templates in GIFT: Analogical Reasoning and Honesty-Humility Traits

Elizabeth Rodriguez[1], Jeanine A. DeFalco[2]
[1]United States Military Academy, West Point, NY, [2] Oak Ridge Associated Universities, USA/US Army Combat Capabilities Development Command Soldier Center – Simulation and Training Technology Center, USA

INTRODUCTION

Tactical Combat Casualty Care (TC3) is taught to members of the United States military at all levels to ensure Soldier safety. It is the process of responding to a casualty in the middle of a combat engagement. The United States Army utilizes medical training programs to educate and evaluate TC3 for all Soldiers. According to the Department of Defense, there are currently over one million US Army Soldiers which means that each of those Soldiers have experienced some basic level of TC3 during their Army service (DOD, 2018). Improvement in the education of TC3 is essential to help ensure Soldier safety and begin to lower the 90% of combat related deaths that occur before the injured Soldier reaches higher level medical care (Kotwal, 2011). The Army is also making the shift to more digital and online training to expedite learning and save on costs of hands on training for each Soldier (Army.Mil, 2013). The balance that must be struck in the future of army training is mitigating the expense while ensuring that adequate training is being distributed to all personnel.

Adaptive online educational tools are the way forward for the Army to expedite learning and is an explicit effort of the Army Research Lab's Essential Research Area: "Accelerated Learning for a ready and Responsive Force" (DeFalco, 2018). In the effort of supporting expertise development, Jung (2016) and Hoffman et al., (2013) recommend fostering high-level reasoning skills. According to the Center for Advancement of Learning and Assessment (King, Goodson, & Rohani, N.D.), higher order thinking skills include critical, logical, reflective, metacognitive, and creative thinking, and are activated when individuals encounter unfamiliar problems, uncertainties, questions, or dilemmas.

Within this framework, then, supporting an accelerated learning pathway to develop the cognitive skills of an expert includes supporting the development of creative thinking--specifically creative reasoning-- a core element of cognitive readiness. Further, we conceptualize an accelerated learning pathway as a pedagogical design template that would be used to accelerate learning in an adaptive instruction system (AIS) that would sequence adaption of instruction according to salient learner traits, in this case personality traits. Pedagogical design templates contain specific, ready-to-be-used content and/or information to inform pedagogical decision-making and instruction that may or may not align to specific learning theories but can simply and streamline pedagogical planning and designs (Dobozy & Dalziel, 2016)—a useful tool for supporting transdisciplinary learning in adaptive instructional systems (AISs) such as the Generalized Intelligent Framework for Tutoring system (GIFT).

The first step in developing a pedagogical design template that would support accelerated medical expertise in an AIS includes understanding what learner traits are correlated with analogical and creative reasoning. Accordingly, our first experiment sought to determine whether there were significant positive correlations between personality traits as measured by the HEXACO with mental rotation tasks and analogical reasoning tasks—two approaches to measure an individual's creative and analogical reasoning skills. This paper reports on our initial findings derived from our first correlational study conducted at the United States Military Academy in the fall of 2018.

ANALOGICAL REASONING AND DECISION-MAKING

Cognitive tasks

The importance of analogical reasoning on decision-making is well evidenced in prior related research. For example, Breuning (2003) looked at the impact that analogical reasoning had on foreign policy decision-making. Analogical reasoning in foreign policy is related to historical case-based explanations to determine how to interact with the current political climate (Breuning, 2003). His study analyzed a 1950s Senate Hearing transcript by breaking each paragraph into either case-based, explanation-based, or model-based reasoning. Breuning (2003) found that approximately 75% of the speakers at the hearing used more explanation-based reasoning than analogical (case-based). The results can be generalized to cognitive models in determining some of the ways that people think and make decisions. The cognitive processes involved in analogical reasoning may seem confusing for foreign policy, but it serves as a spring board for how people remember and react to evolving information.

Additionally, Visser's (1996) identified the two different types of "spontaneous" use of analogy in design. The study analyzes the question of ill-defined problem solving from a cognitive psychology viewpoint by looking at action-execution and action-management analogies. Visser (1996) also explains that there is a gap in the literature on analogous sources and their carry over to tasks in the real world. Action execution is more related to developing and expelling the solution to a complex problem whereas action management looks to accomplish the next action that needs to be executed (Visser, 1996). The participants were recorded performing mechanical tasks, and the results showed that the greater the distance between the target task and the source task, the greater difficulty in creating the analogy. The integration of various types of analogical reasoning can be applied to intelligent tutoring software by helping the developer of the training tap into the ways that people think about learning and solving problems.

In addition to analogical reasoning, one's skill in mental rotation tasks is another competency related to decision-making. Ganis and Kievit (2015) claimed that mental rotation tasks are one of the most influential paradigms in the history of cognitive psychology. Three-dimensional software was employed by Ganis and Kievit to generate 384 objects for rotation with both a baseline object and a target object. Importantly. Mental rotation can predict performance variables such as surgical and spatial skills. 54 participants (31 females) were tested individually at about 60cm from a computer screen and carried out two blocks of 48 trials. The results displayed a linear increase in response time and error rates with angular disparity (Ganis & Kievit, 2015). This means that the greater disparity, the larger the response time from the participants.

Further, Lufler, Zumwalt, Romney and Hoagland (2011) studied the correlational relationship between anatomy student's performance in the course and their visual-spatial ability. 352 first year medical students completed the Mental Rotations Test before the gross anatomy course and 255 at the completion of the course in 2008 and 2009 (Lufler, Zumwalt, Romney & Hoagland, 2011). They determined that students who scored in the highest quartile of the MRT were 2.2 times more likely to sore over 90% on the practical examinations and on both practical and written exams (Lufler, et al., 2011). This is a significant connection to GIFT's application for TC3 because if Soldiers can consistently do well with the mental rotation tasks then they arguably will have a greater likelihood of increased performance in real-world application. While mental rotation is important for creativity, it must be coupled with the ability to learn quickly and under pressure to have a positive impact on Soldier training.

Accelerated learning

Accelerated learning is a strong driving force behind the ideas of GIFT and other digital learning platforms in that they hope to educate the learner effectively and efficiently. Accelerated learning is defined by Hoffman, Feltovich, Fiore, Klein and Ziebell (2009) as not only the hastening of basic proficiency in a task but also encompasses the achievement of expertise. Cognitive flexibility and transformations can help to explain how people can react to accelerated learning. Flexibility is a person's ability to understand their own mental barriers to learning and determine the way around that block (Hoffman, Feltovich, Fiore, Klein, & Ziebell, 2009). Transformation refers to

the necessity of unlearning a task in order to eventually become an expert in that area. Hoffman explains that these factors, combined with a supportive mentor and corrective feedback, help to facilitate successful accelerated learning.

Hoffman (2010) provides a summarized framework for the elements within accelerated learning. He uses the military as an example of where this problem is constantly arising. There are two different forms for transferring knowledge with an accelerated learning framework: transfer across mission types and transfer across responsibility (Hoffman, 2010). In the military, this transfer happens constantly through the changing responsibilities from deployments to changing jobs throughout the military service time. Transfer across missions refers most directly to a Soldier's doctrinal knowledge of the different types of missions they will conduct. Each individual Soldier knows the difference in mission between an ambush and movement to contact. Transfer across responsibilities is similar to the role of a squad leader who is promoted to platoon sergeant. That Soldier needs to know what a good squad leader does and how to coordinate an entire platoon as the highest ranking Non-Commissioned Officer.

Personality and Academic Performance

In a review of the literature, Batey and Furnham (2006) note that while creativity in terms of the production of ideas is related to intelligence, creativity as originality rests largely on personality factors. O'Connor and Paunonen (2007) studied the relationship between the Big Five personality traits and post-secondary academic achievement. This review of other studies uncovered that Openness to Experience was found to be positively correlated with scholastic achievement while Extraversion was negatively correlated (O'Connor & Paunonen, 2007). The current research on the Big Five lends itself to the importance of identifying the types of learners to potentially develop curriculums to improve levels of academic performance in the future (O'Connor & Paunonen, 2007).

Chamorro-Premuzic and Furnham (2008) conducted a study that analyzed the relationship between the personality traits of Openness, Conscientiousness, and cognitive ability and learning. Ability was measured by the Baddeley Reasoning test of fluid intelligence (*gf)* and the Wonderlic Personnel Test IQ (Chamorro-Premuzic & Furnham, 2008). The experimenters defined the learning levels as either surface, deep, or achieving and then had the students conduct four tests by then end of their first month at the university and again during their second year (Chamorro-Premuzic & Furnham, 2008). The results showed that exam marks were significantly correlated with the three personality traits tested. Specifically, Openness had a high positive correlation with IQ, and IQ was strongly correlated with academic performance (Chamorro-Premuzic & Furnham, 2008).

In an additional study, Chamorrow-Premuzic and Furnham (2009) hypothesized that Openness to Experience would have a positive relationship with deep learning. They tested 852 students on the Neuroticism- Extraversion-Openness- Five Factor Inventory (NEO-FFI), as well as a 42-item questionnaire that focused on the reasoning behind how students learn (Chamorro-Premuzic & Furnham, 2009). The results were binned into surface, deep, and achieving categories, showing that Openness to Experience and deep learning were positively correlated.

However, while there is a more robust body of evidence that employs the Big Five (or five factor model) as it relates to intelligence, we have made the choice to employ the HEXACO personality instrument (Ashton & Lee, 2007) as it includes a six trait—Honesty-Humility—that we hypothesize is implicated in positive learning outcomes. Importantly, the Honesty-Humility factor out predicted all factors of the Big Five for correlations with respect to an overt integrity test and business ethical dilemmas task (Ashton & Lee 2007). This is incredibly important to both college and military training tools in that a high Honesty-Humility score can be predictive of a decreased likelihood to cheat others. Also, Openness to Experience reflected in participants an increased opportunity for gains from the energy and time spend in the areas that the participant was interested in (Ashton & Lee, 2007).

CORRELATIONAL STUDY

Research questions and hypotheses

The overarching research question for this work seeks to determine the whether there are statistically significant correlations between analogical/creative reasoning tasks and spatial reasoning tasks with the personality traits measured by the HEXACO and the Short-Item Grit Scale (Duckworth & Quinn, 2009).

The first hypothesis maintained that there would be a statistically significant correlation between the HEXACO personality factor of Openness to Experience and a subject's performance on the analogical reasoning and mental rotation tasks. The second hypothesis stated there would be a statistically significant relationship between the HEXACO personality factor of Honesty-Humility and a subject's performance on the analogical reasoning tasks. The third hypothesis stated there would be a statistically significant positive correlation between the mental rotation and analogical reasoning tasks.

Participants

128 participants (m= 19.66 SD = 1.464) and then received 5 points of extra credit for their introductory psychology class for their participation in the study. In the study, 23 participants self-identified as a novice, 73 self-identified as a journeyman, and 1 self-identified as an expert in the field of combat casualty care treatment. The cadets make up a diverse population of 18-22 years old from across the United States and some allied nations.

Apparatus

The original plan for running this correlational study was to use GIFT to deliver the assessment instruments. However, at the time this study was ready to be launched, it was discovered that GIFT did not have the capability to design a timed question that would launch subsequent questions at the time expiration. For the mental rotation tasks, to obtain a more accurate measurement of a person's spatial ability, the instrument is designed so that participants only have 7.2 seconds to respond whether the images are the same or different before loading the next image. At the time of writing this paper, timed questions have now been added as a functionality into GIFT, but this functionality was not integrated at the time of running this first correlational study. With that limitation in mind, this experiment utilized Qualtrics to distribute the survey and the SONA system at USMA to obtain participants and provide those participants with extra credit points. Qualtrics is an online survey software that allows the experimenter to digitally upload their survey for participants to complete. The survey consisted of the demographic questionnaire, the short item grit survey (Duckworth & Quinn, 2009), the HEXACO personality test (Ashton & Lee, 2007), the Analogical Finding Task Matrix (AFTM) (Weinberger, Iyer, & Green, 2016), and mental rotation tasks (Ganis & Kievit, 2015).

The Short-Item Grit survey determines how the attribute of Grit supports or impedes creative reasoning (DeFalco, 2018). The eight grit questions are scored on a five-point Likert scale ranging from "not like me at all" to "very much like me" in response to questions like "I am a hard worker" and "setbacks don't discourage me" (Duckworth & Quinn, 2009). Participants were not limited in their time to answer these questions, but the average response time was five minutes to complete the survey.

The HEXACO personality survey took approximately twelve minutes to complete 60 questions, where participants respond on a five-point Likert scale.

The analogical task finding matrixes (there were two) asked participants to match 10 analogical pairs with one other pair with only one response per each analogical pair. This portion of the survey took approximately eight minutes. For example, participants could be given "Watermelon/Rind" and they could match it with "Orange/Peel."

The mental rotation tasks took 7.2 seconds each and there was a total of 40 pairs delivered. In this task, the participant must identify if the three-dimensional shapes are the same or different before another pair of images are displayed for a response.

Research design

The research design was a correlational study to determine if there was a statistically significant positive relationship between the analogical reasoning and mental rotation tasks and any of the traits measured by the HEXACO, as well as to determine the relationship between the HEXACO traits and Grit, as well as the mental rotation tasks and analogical reasoning tasks.

Procedure

For the correlational study, participants were informed of potential extra credit survey opportunities through the SONA system and could access the survey either on a mobile device or on their laptops. The participants could choose from a list of approved surveys and 128 of them selected "Developing Accelerated Learning Models in GIFT." The consent form was the first screen to appear to the participants. After consenting to the experiment, participants answered the twelve demographic questions for approximately five minutes.

Next, participants responded to the ATFM and matched two sets of 10 analogical pairs for roughly eight minutes. The participants then answered either "same" or "different" for the 40 mental rotation tasks, with a forced response time to occur within 7.2 seconds for each pair of shapes. Then the participants took the Short-Item Grit survey and then the 60 HEXACO questions for ten minutes. Participants had to complete all questions from the previous section before moving on to the next set of questions. At the completion of the Grit and HEXACO portion, participants had completed the entire survey and experimenters were able to assign extra credit points for their participation. The total time for this study was roughly 35 minutes.

Results

Descriptive data is displayed in Table 1 below. The discrepancy in sample sizes for the different portions of experiment one was likely caused by internet connection issues and the various time schedules that cadets at USMA operate under.

Table 1. Descriptive Statistics of Test Subjects

Variable	Statistic 1	Statistic 2	Statistic 3
Age (*N*=92)	Mean = 19.66	*SD* = 1.464	Range: 18-24
Gender (*N*=92)	Female = 33	Male = 64	
Military Service (*N*=92)	Yes = 15	No = 82	
Critical Care Knowledge (*N*=92)	Novice = 23	Journeyman = 73	Expert = 1
Mental Rotation Score *(N*=92)	Mean = 32.34	*SD* = 6.288	Total Score = 40
Analogical Reasoning Semantic Distances: Matrix 1 (*N*=97)	Mean = 774.38	*SD* = 274.251	Total Score = 1069
Analogical Reasoning Semantic Distances: Matrix 2 (*N*=97)	Mean = 603.91	*SD* = 298.52	Total Score = 1052
Openness to Experience (*N*=92)	Mean = 32.63	*SD* = 6.326	

After data collection from Qualtrics, data was cleaned and analyzed, correlational analyses were run in SPSS.

For the analogical reasoning tasks there were two matrices used to measure both creativity and analogical reasoning. The semantic distance score for the analogical reasoning tasks gives the measure of strength of an individual's reasoning level (DeFalco, 2018). There was a statistically significant correlation between the Openness to Experience score on the HEXACO traits and the semantic distance score of the analogical reasoning task (r= 0.279, N= 92, p= 0.007). There was also a positive relationship between the analogical reasoning task and the mental rotation tasks (r=0.444, N=95, p=0.000).

Also, by splitting the groups into high (>33) and low (<32) of the Honesty-Humility factor, there was a statistically significant difference in the means sematic distance of analogical reasoning tasks in the first matrix, $F(2,94) = 7.046$, $p = 0.001$. There was also a positive correlation between GRIT and HONESTY-HUMILITY, r = .343, $N = 92$, p= 0.001. There was a positive correlation between the score of the semantic distance of analogical reasoning tasks in the second matrix and the HONEST- HUMILITY score, r= 0.332, $N = 92$, $p = 0.001$.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

The results of this correlational study confirmed that there was a statistically significant positive correlation between Honesty-Humility and an individual's creativity levels. Also, there was a statistically significant positive correlation between Openness to Experience and a subject's performance on the analogical reasoning tasks. This information is relevant for informing the design of a future experiment that will determine whether the sequencing of content with analogical/creative reasoning tasks contributes to an acceleration of medical decision-making expertise within the domain of critical care. With this data, we expect to make significant strides towards validating a transdisciplinary pedagogical design template that can become part of the suite of tools integrated into the dashboard of GIFT's authoring tools.

REFERENCES

Artino, A., Army.mil. (2013) USAPHC Web-based courses save money, satisfy demand. Retrieved from https://www.army.mil/article/94773/usaphc_web_based_courses_save_money_satisfy_demand

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11(2), 150-166. doi:10.1177/1088868306294907.

Batey, M., & Furnham, A. (2006). Creativity, intelligence, and personality: A critical review of the scattered literature. *Genetic, social, and general psychology monographs*, *132*(4), 355-429.

Breuning, M (2003). The role of analogies and abstract reasoning in decision-making: evidence from debate over Truman's proposal for development assistance. *International Studies Quarterly*, (2), 229.

Chamorro-Premuzic, T., & Furnham, A. (2008). Personality, intelligence and approaches to learning as predictors of academic performance. *Personality and Individual Differences*, *44*(7), 1596–1603.

Chamorro-Premuzic, T., & Furnham, A. (2009). Mainly Openness: The relationship between the Big Five personality traits and learning approaches. *Learning and Individual Differences, 19*(4), 524-529. doi:10.1016/j.lindif.2009.06.004

DeFalco, J. (2018). *Human Subjects Research Protocol*. Unpublished manuscript along with statistic assistance, United States Military Academy, West Point.

DOD. (2018). Legacy Homepage. Retrieved from https://dod.defense.gov/News/Special-Reports/0518_budget/ Dobozy, E., & Dalziel, J. (2016). Transdisciplinary Pedagogical Templates and Their Potential for Adaptive Reuse. *Journal of Interactive Media in Education*, *2016*(1).

Duckworth, AL, & Quinn, P.D. (2009). Development and validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment*, 91, 166-174. http://www.sas.upenn.edu/~duckwort/images/Duckworth%20and%20Quinn.pdf

Ganis, G., & Kievit, R. (2015). A new set of three-dimensional shapes for investigating mental rotation processes: Validation data and stimulus set. *Journal of Open Psychology Data*. 3(1): e3, DOI: http://dx.doi.org/10.5334/jopd.ai.

Hoffman, R. R., Anders, D., Fiore, S. M., Goldberg, S., Andre, T., Freeman, J.,& … Klein, G. (2010). Accelerated Learning: Prospects, Issues and Applications. Proceedings of The Human Factors and Ergonomics Society Annual Meeting, 54(4), 399.

Hoffman, R. R., Feltovich, P. J., Fiore, S. M., Klein, G., & Ziebell, D. (2009). Accelerated Learning. *IEEE Intelligent Systems*, *24*(2), 18–22. https://doi.org/10.1109/MIS.2009.21

Hoffman, R. R., Ward, P., Feltovich, P. J., DiBello, L., Fiore, S. M., & Andrews, D. H. (2013). Accelerated learning: Training for high proficiency in a complex world. New York: Psychology Press.

Jung, E. (2016). *Expertise development through accelerated learning: A multiple-case study on instructional principles* (Doctoral dissertation, Indiana University).

King, FJ., Goodson, L, & Rohani, F. (n.d.) "Executive Summary: Definition." n.d. Higher Order Thinking. 22 October 2011 <http://www.cala.fsu.edu/files/higher_order_thinking_skills.pdf>.

Kotwal, R. S., Montgomery, H. R., Kotwal, B. M., Champion, H. R., Butler, F. K., Mabry, R. L., . . . Holcomb, J. B. (2011). Eliminating Preventable Death on the Battlefield. *Archives of Surgery,146*(12), 1350- 1358. doi:10.1001/archsurg.2011.213.

Lufler R., Zumwalt A., Romney C., & Hoagland T. (2012). Effect of visual–spatial ability on medical students' performance in a gross anatomy course. *Anat Sci Educ* 5:3–9.

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, *43*(5), 971–990.

Visser, W. (1996). Two functions of analogical reasoning in design. A cognitive-psychology approach. *Design studies*, *17*(4), 417-434.

Weinberger, A. B., Iyer, H., & Green, A. E. (2016). Conscious augmentation of creative state

ACKNOWLEDGEMENTS

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-17-2-0152. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

ABOUT THE AUTHORS

CDT Elizabeth Rodriguez *is a senior at the United States Military Academy. She is from St. Cloud, Florida and is working to complete a Bachelor of Science degree in Engineering Psychology. Through the Negotiations Project, CDT Rodriguez gained a 40-hour certification in Crisis Negotiations from the FBI and has been the cadet in charge of the program from 2017-2019. CDT Rodriguez will be branching transportation upon graduation from West Point.*

Jeanine A. DeFalco, PhD, *is currently a Post-Doctoral Research Fellow with the CCDC Soldier Center-STTC working out of the United States Military Academy, in the Department of Behavioral Sciences and Leadership. Her current research includes examining the relationship of creative reasoning on expert problem solving in medical education, mediated by GIFT, and assisting in the design of human virtual agents to support moral and ethical leadership for military personnel. Her PhD is in Psychology, Human Development/Cognitive Studies with a concentration in Intelligent Technologies, from Columbia University.*

THEME V:
GIFT AND FUTURE TRAINING & EDUCATION CONCEPTS

Teamwork Training Architecture, Scenarios, and Measures in GIFT

Robert McCormack[1], Tara Kilcullen[1], Anne M. Sinatra[2], Alexander Case[1], Daniel Howard[1]
Aptima, Inc.[1], U.S. Army Combat Capabilities Development Command (CCDC) Solider Center - Simulation & Training Technology Center (STTC)[2]

INTRODUCTION

The success or failure of teams facing complex tasks often hinges on their ability to communicate and collaborate throughout the mission. Indeed, numerous studies have shown that developing and maintaining strong teamwork skills through training interventions have a positive impact on performance outcomes (Sottilare, et. al. 2017; Wilson, et. al. 2007; Salas, et. al. 2008). The U.S. Army has specifically recognized the impact of teamwork on mission command and the need for commanders to build teams rather than relying on pre-established relationships. Furthermore, effective teams are those that can establish a high degree of coordination both within their unit and across higher, lower, and adjacent echelons (ARDP 6-0, 2012). Training teams to improve their teamwork skills, such as coordination, however, is often treated as an implicit by-product of task-focused training rather than a goal in and of itself. As such, the ability for trainers to explicitly measure and intervene in teamwork skill development is often lacking. Intelligent Tutoring System (ITS) frameworks, such as the Generalized Intelligent Framework for Tutoring (GIFT; Sottilare et. al., 2012; Sottilare et. al., 2017), offer a great deal of potential for team training, but have generally been developed with individual training in mind. In a previous GIFTSym paper (McCormack, et. al., 2018), the authors presented a set of proposed scenarios and an approach to developing teamwork measures within a virtual training environment. This paper builds upon that work and describes an effort to extend GIFT to include the necessary components of teamwork training. We discuss a GIFT architecture that provides team-level skill feedback, while minimizing the infrastructure cost of developing and maintaining Domain Knowledge Files (DKFs). In addition, we present a number of realistic, doctrinally-relevant scenario vignettes developed to provide multiple opportunities for teamwork skill development and feedback. Finally, we provide examples of measures of coordination tailored to the scenario that enable analysis and feedback on skill development.

While the ultimate goal is to enable GIFT to support teamwork training across multiple virtual and live environments, the initial focus of this effort was on training Army teams within Virtual Battlespace (VBS) 3.0. This virtual training environment (VTE) offers multiple advantages: it provides a realistic backdrop for doctrinally-relevant tasks; it enables scalable team sizes and multiplayer experiences; and, much of the infrastructure for enabling integration of VBS and GIFT has already been developed.

Team Training Architecture

In order to instantiate teamwork measures in GIFT, we developed an architecture that would enable measurement across multiple VBS players, without inducing a heavy burden associated with creating and man- aging multiple DKFs. Previous efforts at teamwork measurement required a DKF for each player and additional DKFs for each combination of players and the team as whole (Bonner, et.al. 2017). Any measurements for combinations of dyads, triads, etc. of individuals required a separate DKF. While this allows player and sub-team specific measurement, as the team size grows, the exponential growth of DKFs required by this approach quickly becomes unmanageable. Because our focus is on team-level measures, we chose an architecture that was best suited for that, but still allows future inclusion of individual measures.

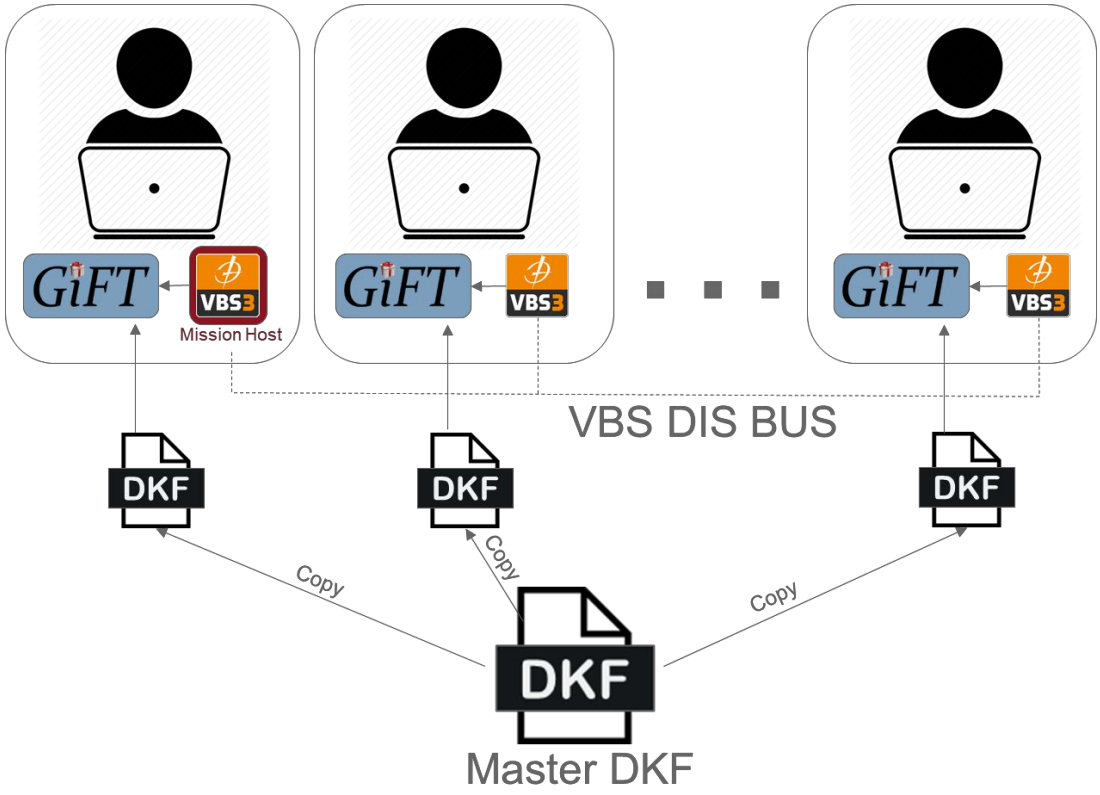


Figure 1. Team Training Architecture and DKF use

Figure 1 provides an overview of the team training architecture. Each team member runs a local GIFT instance as well as VBS. One team member serves as the VBS mission host (this can be any team member, although in practice the team lead would normally take this role) and each other VBS client joins the mission remotely. Our work to date requires manual selection of team roles by each member, but this could be automated in future iterations to automatically assign preset member roles. The VBS DIS (Distributed Interactive Simulation) data bus enables communication between all the client machines and allows the VBS clients to exchange information. In our architecture, this is the only communication occurring between client machines. That is, the GIFT instances themselves do not directly communicate to each other and do not exchange any data or measurements. Instead, they each independently compute measures and feedback based on the information captured from the DIS bus. This teamwork architecture requires development of only a single master DKF that computes all measures at the team level. Each team member's GIFT client runs an instance of GIFT independently with this master DKF. This vastly simplifies the calculation and feedback of teamwork measures.

This approach of a single cloned DKF for all team members has several advantages. The main advantage comes in ease of training content development and deployment. The effort in developing DKFs can be considerable. Use of a single file can reduce the time required for a training content developer to create training packages and allow them to focus on measurement rather than file management. Cloning the GIFT scenarios across all team member machines is fast since there are no individual differences between them. In addition, the single DKF framework ensures all team members receive the same feedback at the same time (ignoring network lag) since all calculations are performed on the information flowing through the DIS bus.

However, there are limitations to this approach. Individualized measurements and feedback are not possible with this current architecture as they would require customization of individual DKFs. We believe, how- ever, that ongoing efforts in improving DKF development and management tools will enable easier customization of DKFs to

individuals in the future. A hybrid approach may be possible where a single team DKF is used in conjunction with individual DKFs to provide both team-level and individual measurement and feedback. Since the effort described here is focused on teamwork measures, the current architecture provides sufficient flexibility at that level.

Team Training Scenario

The scenario developed for this effort consists of a number of distinct, but narratively related, events or vignettes that the team plays through. Although the map follows a mostly linear path, the vignettes are broken apart as independent training elements. This provides the ability to dynamically reconfigure the training to adapt to high or low-performing teams, helps facilitate repeated training so the team is not able to predict how the scenario unfolds, and enables the delivery of training feedback in GIFT at the end of each vignette. In addition, the vignettes each provide a number of opportunities to train a variety of team- work skills. In this section, we describe each of the vignettes, as well as the expected actions and protocols for the team. In the subsequent section, the teamwork measures for these vignettes are discussed.

The scenario is adapted from an Army Basic Leader Course (BLC) Combat Search and Rescue (CSAR) training scenario. The team consists of a nine-member squad: two four-person fire-teams and a squad leader. The scenario takes place along a linear path through a wooded area. In the scenario, an F-16 pilot has ejected and landed in the area and their medical condition is unknown. The primary objective is to perform a search and rescue for the downed pilot, with a secondary objective of reaching a small village after the pilot is rescued. The team receives intelligence that enemy militia are known to be in the area and are hostile to our presence. Figure 2 depicts the scenario map and layout of each of the vignettes.

Vignette 1: Encountering an Improvised Explosive Device (IED)

Within the scenario, there are two vignettes (vignettes one and four) in which the team encounters a potential IED threat. One is a hoax IED and one is real, although the team is not aware of the status of either. It is expected that they treat both as potentially deadly. The IEDs are represented as deceased dogs in the road, a known method of IED camouflage. Wires protruding from the animal provide further indication of the threat. Upon identification of the IEDs, the team member who first notices it is expected to halt and inform the rest of team of the threat. The team then coordinates their sectors of fire to ensure they have full 360-degree situational awareness of the environment. This requires coordination and communication within the team to secure and clear the area. Next, the team is expected to address the 5 “C”s of IEDs. These are Confirm, Clear, Call, Cordon, and Control. These choices are given through a menu system within VBS and the team is expected to choose the right “C”s in the right order.

Vignette 2: Finding and Rescuing the Pilot

As the team enters the area known to contain the downed pilot (the general area is provided in the pre- briefing material), they are expected to maintain formation and begin visually scanning the area. The pilot is located near the path, but hidden amongst the foliage. This requires the team to search the area. If not found in a set amount of time, the pilot will release a flare to draw the team to her location. Once the pilot is located, the team is expected to coordinate a number of activities including securing the pilot, setting up a cordon with interlocking sectors of fire to scan for threats, applying first aid, moving the pilot to an open landing zone, issuing a 9-line MEDEVAC request, and waiting for the helicopter to extract the pilot.

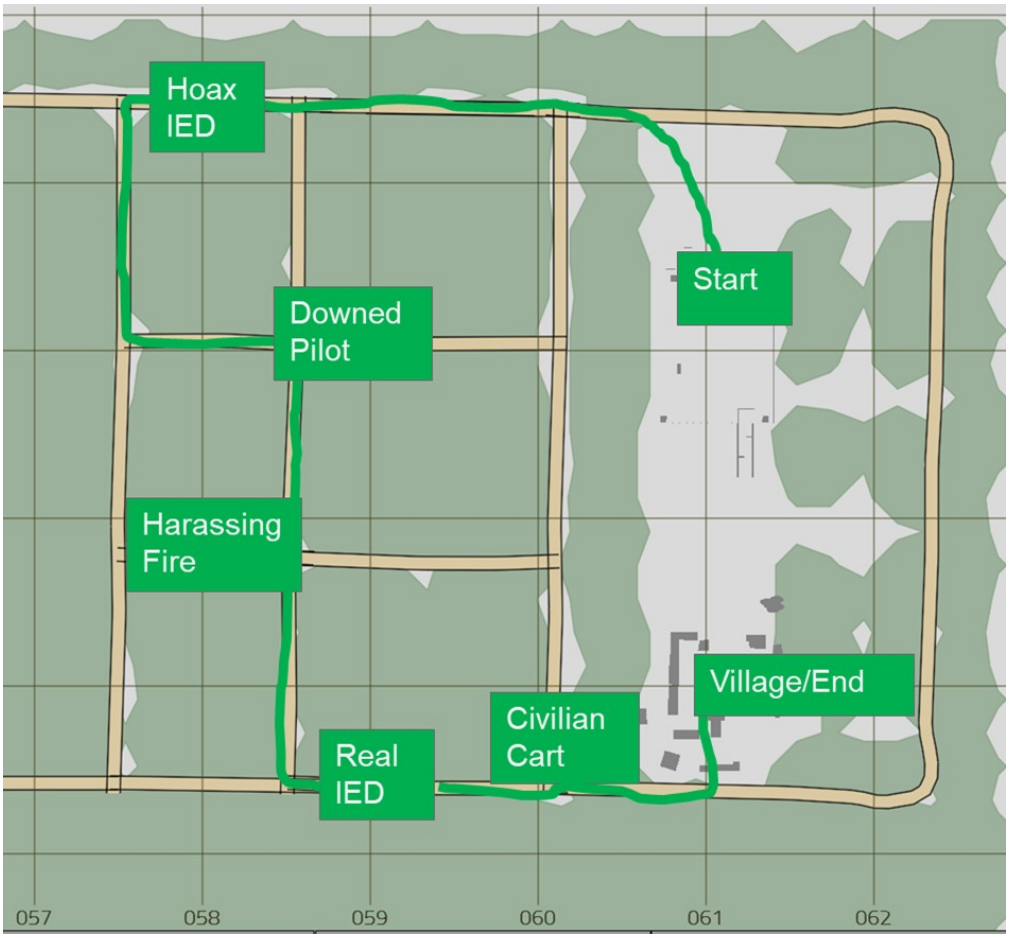


Figure 2. Scenario Map and Vignette Layout

Vignette 3: Reacting to Harassing Fire from Hostile Militia

At a point along the path, the team encounters harassing small-arms fire from two hostile militia members. This occurs suddenly and without warning, requiring the team to react quickly. In VBS the damage and accuracy of the hostile actors is reduced so as not to injure any of the team members (although this fact is not known to the team). The goal of this vignette is for the team to coordinate an immediate response to the threat without jeopardizing the other mission objectives. We require the team to contain the threat by returning fire to either kill or scare-off the hostile actors, but we do not expect them to chase after the militia. Specifically, the team is expected to take cover, return fire, call clear once the threat is neutralized, and continue on the path.

Vignette 4: Encountering an Improvised Explosive Device (IED)

The fourth vignette involves encountering a real IED (although the team does not know whether this one is real or another hoax as in vignette 1). The goals, actions, and measures for this vignette are the same as vignette 1, with the exception that if the team gets too close to the IED, it will detonate, ending the vignette.

Vignette 5: Reacting to Unknown Individual

In the final vignette, the team encounters an individual pulling a cart along the road. The individual's identity and intentions are unknown. The individual is heading directly towards the team, and while not overtly threatening, he still poses a potential risk. The team is expected to quickly notice the individual, halt movement, communicate his

presence and approach to the rest of the team, and "Show, Shout, Shoot," that is, escalate fire at the individual until he retreats or is otherwise neutralized. The objective is for the team to react quickly, communicate the evolving situation to each other, and secure the individual (either through lethal force or until he leaves).

Team Training Measures

Measurement and assessment of teamwork skills within the training environment is crucial to ensure that the system is providing the correct opportunities for skill development, but also to provide the team feed- back on their teamwork training progress. The development of measures for this effort occurred largely in parallel with scenario development. Specific elements of the vignettes were chosen because they provide opportunities to assess teamwork skills. To develop teamwork measures, we use a process based on the Rational Approach to Developing Systems-based Measures (RADSM; Orvis et al., 2013), which has been successfully used to develop indicators and measures of team states (McCormack, Brown, Orvis, Perry, Myers, 2017). The RADSM process consists of several steps that ensure that developed measures are conceptually sound and contextually relevant. The end result of this process is a set of teamwork measures that can be assessed automatically and unobtrusively (that is, not requiring human coding or input) given the data available in the system. We describe this process in our previous GIFTSym paper (McCormack, et. al., 2018), so we focus here only on the resultant measures.

Selection of the specific teamwork construct to train was motivated a previous meta-analysis (Sottilare, et. al., 2017) in which a number of teamwork themes were identified, including coordination, cohesion, communication, cooperation, conflict, and others. The teamwork measures developed for this effort focus primarily on coordination. This construct is defined in various ways throughout literature, but for our purposes we treat coordination as the synchronization and awareness of team member actions in pursuit of a common team goal. Communication among team members is a large part of coordination (and most teamwork constructs), but communications were not able to be captured and analyzed in GIFT during this effort. Verbal communication analysis requires the ability to capture individual speech utterances from each team member and perform speech-to-text processing on the audio. There is currently no approach integrated into GIFT to analyze verbal communications and it was beyond the scope of this effort to develop one. Furthermore, while the text chat (such as instant messaging apps or VBS's built-in capabilities) offers an alternative solution, it was deemed untenable for this scenario. Typing messages would require participants to stop moving and performing actions within VBS, which would interfere with the often highly-kinetic vignettes. As such, our focus of measurement was not on the communications themselves, but on the actions and behaviors that require communication and coordination to occur.

We describe a selection of coordination measures developed for this scenario in Table 1. Several of these measures repeat across vignettes (such as maintaining team formation) and others are omitted here for the sake of brevity.

Table 1. Example Measures of Coordination

Measure Name	Description	Measure Feedback
Maintaining Team Formation	During movement along the road the team should maintain a close formation. This initial measure will focus on the geodesic distance of the team, rather than on the relative positioning of each player. Geodesic distance is defined as the farthest distance any two team members are apart. That is, we measure the distance between each pair of individuals, and take the maximum. The geodesic distance and the teams (above/at/below) expectation rating should be reported every 15 seconds.	Above Expectation: 8m < Geodesic distance < 12m
		At Expectation: 5m < Geodesic distance < 8m OR 12m < Geodesic distance < 15m
		Below Expectation: Geodesic distance < 5m OR Geodesic distance > 15m
Completing a Team Halt	The first person to notice a threat (IED, unknown individual in the road) should halt and give the command to teammates to halt. We define a maximum and minimum distance from the threat. Once an individual is within the maximum distance radius, we monitor their movement and look for a stop of movement. We then monitor other team members to identify if/when they stop movement. If an individual has reached the mini- mum radius before a halt occurs, the entire team fails.	Above Expectation: first person halts between mini-mum/maximum distance, calls halt, other team members halt within 3 seconds of call.
		At Expectation: first person halts between minimum/maxi-mum distance, calls halt, some team members halt within 3 seconds, but other team members halt before reaching mini- mum distance, but after 3 seconds.
		Below Expectation: Any team member crosses minimum distance radius before halt is complete.
Completing the 5 C's of an IED Encounter	After noticing an IED, the team correctly selects the "5 C's" (Confirm, Clear, Call, Cordon, Control) from the drop down menu in the right order. Each person may only select one item from the list, and the selection should be completed in a timely manner. Team members should communicate and coordinate on who is completing a selection and the correct order of se-lection.	Above Expectation: 5 different individuals correctly choose the 5 "C" s in the right order from the menu. Completes se-lection within one minute.
		At Expectation: At least 4 different individuals choose the 5 "C"s, but in the wrong order or it takes longer than one minute, but less than two minutes.
		Below Expectation: The team fails to select the correct 5 "C"s, makes an incorrect selection, or takes longer than two minutes, or selections are only made by 3 or fewer individuals.
Attaining Visual Control of the Environment	The team should scan their surroundings for potential threats. To measure this, we take each team members orientation angle and assume that they can visually assess a +-30-degree arc from their orientation angle. The union of all of these field of view (FOV) arcs across the team should cover 360 degrees.	Above Expectation: There are less than 10 degrees not covered in the union of FOVs.
		At Expectation: There are between 10 and 30 degrees not covered in the union of FOVs.
		Below Expectation: There are greater than 30 degrees not covered in the union of FOVs.

Table 1 (Continued). Example Measures of Coordination

Measure Name	Description	Measure Feedback
Identifying Hostile Actor Location	After the first hostile bullet is fired we measure the time it takes to identify the location of the hostile actors. The timer ends when one of the criteria is fulfilled: at least one team member has spotted the hostile actors through the binoculars or at least half the team members have their weapons pointed within +-10-degrees of the hostile actors.	Above Expectation: The team spots the hostile actors within 10 seconds.
		At Expectation: The team spots the hostile actors within 20 seconds.
		Below Expectation: The team takes longer than 20 seconds to spot the hostile actors.
Applying First Aid	After finding the pilot, the team correctly selects the MEDEVAC actions (Assess the pilot's condition, Identify and Control Bleeding, Assess Breathing and Chest Injuries, Check for Burns, Monitor for Shock) from the drop down menu in the right order. Each person may only select one item from the list, and the selection should be completed in a timely manner. Team members should communicate and coordinate on who is completing a selection and which is the correct order of selection.	Above Expectation: 5 different individuals correctly choose the First Aid steps in the right order from the menu. Completes selection within one minute.
		At Expectation: At least 4 different individuals choose the First Aid steps, but in the wrong order or it takes longer than one minute, but less than two minutes.
		Below Expectation: The team fails to select the correct First Aid steps, makes an incorrect selection, or takes longer than two minutes, or selections are only made by 3 or fewer individuals.
Taking Cover from Harassing Fire	After the first bullet is fired by the militants, we measure the amount of time it takes for all team members to take cover. In game, we define taking cover as either lying prone or crouching. The outcome metric is the amount of time between the first hostile bullet fire and the last team member to take cover.	Above Expectation: The team takes cover within 5 seconds.
		At Expectation: The team takes cover within 10 seconds.
		Below Expectation: The team takes longer than 10 seconds to take cover.

While the measures discussed here were initially developed with VBS in mind, we note that there is high potential for transfer to other learning environments. By abstracting away the specific doctrinal details, we arrive at general teamwork measures and GIFT condition classes that can be instantiated elsewhere. As an example, consider the "5 C's" measure from the IED vignette and the first aid measure from the pilot rescue vignette. We refer to these as "team sequence" tasks, where team members must select the right choices, in the right order, across different team members. The developed condition class for this is applicable to any training environment where coordinated selection of ordered choices is required. Another example is the team formation measure, where team members are required to stay relatively close but not bunch up. The corresponding condition class measures the geodesic distance of the entire team and compares that against acceptable thresholds. This condition class is applicable to any spatially-oriented training environment, where relative location of team members is important.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

The new architecture for training teams in GIFT discussed in this paper provides a straightforward, scalable approach to delivering teamwork skill training content. By utilizing one team-level DKF, content develop- ers can focus on measure development, implementation, and training delivery. While currently only team- level feedback can be delivered through this approach, individualized or sub-team feedback will be possible with future implementations.

The scenario developed under this architecture provides realistic team training in a virtual environment. The individual vignettes are reconfigurable and provide opportunities to develop a variety of teamwork skills. The initial focus of team training has been on coordination, but future work will expand to measure cohesion, communications, conflict management, and others. The new condition classes implemented in GIFT are generalizable to a variety of training scenarios and training content.

Future technology development will also seek to incorporate naturalistic communication assessment through speech-to-text and natural language processing capabilities. This will further enable teams to gain invaluable skills, while minimizing the need for more obtrusive or burdensome communication assessment techniques, such as chat or menu selections, which can distract from the central goals of the training. Finally, additional customizations to GIFT will provide better immediate and after action review feedback through the use of the learner action panel.

ACKNOWLEDGEMENTS

The research reported in this document was performed in connection with contract number W911NF-17- C-0061 with the U.S. Army Contracting Command - Aberdeen Proving Ground (ACC-APG). The views and conclusions contained in this document/presentation are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of ACC-APG, U.S. Army Research Laboratory or the U.S. Government unless so designated by other authorized documents. Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

Army Doctrine Reference Publication (ADRP) 6-0, Mission Command. Washington, DC: Headquarters, Department of the Army, 2012.

Bonner, D., Gilbert, S., Winer, E., Dorneich, M., MacAllister, A., Kohl, A., et al. (2017). Military Team Training Utilizing GIFT. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL.

McCormack, R.K., Brown, T.A., Orvis, K.L., Perry, S., & Myers, C. (2017). Measuring Team Performance and Co-ordination in a Mixed Human-Synthetic Team Training Environment. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL.

McCormack, R., Kilcullen, T., Sinatra, A., Brown, T., & Beaubien, J. (2018). Scenarios for Training Teamwork Skills in Virtual Environments with GIFT. Paper presented at the Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym6), Orlando, FL.

Orvis, K. L., Duchon, A., & DeCostanza, A. (2013, January). Developing Systems-based Performance Measures: A Rational Approach. In The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC) (Vol. 2013, No. 1). National Training Systems Association

Salas, E., Cooke, N.J., & Rosen, M.A. (2008). On teams, teamwork, and team performance: Discoveries and developments. *Human Factors, 50*, 540-547.

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

Sottilare, R. A., Burke, C. S., Salas, E., Sinatra, A. M., Johnston, J. H., & Gilbert, S. B. (2017). Designing adaptive instruction for teams: A meta-analysis. *International Journal of Artificial Intelligence in Education*, 1-40.

Wilson, K. A., Salas, E., Priest, H. A., & Andrews, D. (2007). Errors in the heat of battle: Taking a closer look at shared cognition breakdowns through teamwork. Human Factors, 49, 243–256

ABOUT THE AUTHORS

Dr. Robert K. McCormack *is a Principal Mathematician and Deputy Director of the Performance Assessment Technologies Division at Aptima. He has expertise in the areas of unobtrusive measurement, computational linguistics (NLP), machine learning, epidemiological modeling, and human sociocultural modeling and analysis. Dr. McCormack received a Ph.D. and M.S. in Mathematics from Texas Tech University, and a B.A. in Mathematics and Computer Science from Austin College*

Ms. Tara Kilcullen *is a Program and Customer Engagement Lead at Aptima, Inc. In her role, she supports business strategy, planning, and execution, defines market needs for technology requirements, product maturation and successful transition of science and technology (S&T) research programs, and leads defense programs that differentiate Aptima as an industry leader in Modeling, Simulation and Training. Ms. Kilcullen holds B.A.'s from the University of Pittsburgh and an A.S. from Full Sail University as well as several certifications.*

Dr. Anne M. Sinatra *is a Research Psychologist, and part of the adaptive training research team within CCDC Soldier Center – STTC's Learning in Intelligent Tutoring Environments (LITE) Lab. She works on the Generalized Intelligent Framework for Tutoring (GIFT) project. Her background is in Human Factors and Cognitive Psychology.*

Mr. Alexander Case *is a Software Engineer in the Product Engineering Division at Aptima, Inc. His technical interests focus on front- and back-end applications, data collection and analysis, and video game and simulation develop- ment. Other areas of interest include application telemetry, containerization, continuous integration, and configuration management. While at Aptima, he has worked on Angular web applications and developed server-side applications in Python and C#. Mr. Case holds a Bachelor's degree in Computer Science from Bridgewater State University.*

Mr. Daniel Howard *is a Principal Software Engineer and Product Line Lead for Data Management and Visualizations at Aptima, Inc. Mr. Howard works on developing a variety of applications focusing on complex visualization and the analysis of large volumes of data, and is responsible for the efficient storage and retrieval of this measurement data through Aptima's ASA™ framework. Mr. Howard holds a M.S. and B.S. in Computer Engineering from Rochester Institute of Technology.*

Authoring Team Tutors in GIFT: An Automated Tool for Alignment of Content to Learning Objectives

Benjamin Bell[1], Keith Brawner[2], Elliot Robson[1], Debbie Brown[1], Elaine Kelsey[1]
Eduworks[1], US Army RDECOM-NSRDEC[2]

INTRODUCTION

The process of authoring Intelligent Tutoring Systems (ITS) in the Generalized Intelligent Framework for Tutoring (GIFT) (Sotillare et al, 2013) is assisted by a continually-evolving collection of authoring tools. These tools can accelerate GIFT development by supporting instructional design tasks like sequencing, feedback, adaptation, and assessment. With growing demand for team tutoring in support of rapidly- evolving Army requirements, GIFT tutors must be able to scale learning to meet team training needs, be capable of incorporating broad content; and offer instructional value for both individual Soldiers and teams (Sottilare et al, 2011; Sottilare et al, 2018; Salas et al, 2015; Sottilare et al, 2017b; Fletcher & Sottilare, 2017). A key need is to help ITS authors efficiently find and maintain relevant content, and to assist authors with discriminating between content supporting individual learning objectives and team learning objectives. Addressing this need efficiently calls for automation that supports the analysis of information and its alignment with learning objectives (LOs) (Bonner et al, 2016).

In this paper we introduce a new authoring aid, to be incorporated within GIFT, to help ITS developers find, organize, and curate resources aligned with desired individual and team learning objectives. ***Ma- chine-Assisted Generation of Instructional Content*** **(MAGIC)** analyzes source documents and extracts content that aligns with specified learning objectives. MAGIC additionally lends much-needed support for team training development by performing this alignment for both individual and team learning objectives. Building on and extending existing artificial intelligence (AI) and natural language processing (NLP) techniques, MAGIC will streamline content alignment, distinguish between individual and team content, and help extend the reach of GIFT tutoring to meet Army team training demands.

MAGIC will contain three layers: backend algorithms and analytics, services and APIs for integration into the GIFT ecosystem, and integrated end user tools for authors. In this paper we describe our initial focus on developing the underlying techniques used by MAGIC and creating a prototype interface for training developers that presents algorithmic outputs. We will describe our work in extending existing NLP and machine learning (ML) libraries to extract and organize learning objectives and integration of these libraries to create an LO repository. A novel aspect of this work is applying ML models to the discrimination between individual and team LOs. We then describe development of automated methods for aligning excerpts of content with LOs and specific roles within a team.

Our plans include implementing an end-user toolset for integration with GIFT to help authors organize LOs and topics and tag, find, sort, and repurpose content that aligns with given LOs and role-based parameters. Finally, we discuss our longer-term plans for incorporating multimedia resources by applying automated transcription techniques. The work we present will advance the state-of-the-art in applying machine learning and NLP to authoring and development of training and in particular team tutoring, and will extend GIFT by supporting authors in collecting and aligning content with individual and team learning objectives. (Gilbert et al, 2017; McCormack at al, 2018; Sotillare et al, 2017a; Sinatra, 2018)

SCALING TEAM TRAINING

Scaling virtual training for teams to fully address Army needs requires tools and techniques for efficiently creating team tutoring simulations. While GIFT supports several instructional design tasks, finding and organizing content that aligns with desired learning objectives remains a labor-intensive process that takes place outside of GIFT. Achieving scale means that virtual training must span broad content. Maintaining relevance means that virtual training must be readily adaptable as learning needs shift in response to equipment upgrades, changes in tactics, evolving threats, and operations in new theaters.

To benefit teams, authors of team tutoring must navigate complicated content management tasks related to distinguishing content that supports individual skills and content aligned with team skills, as well as trying to identify content associated with specific roles within a team. Creating and maintaining virtual team training systems thus remains costly and time-consuming. To help developers of team training find and tag relevant content more efficiently, automation is needed that supports analysis of content and its alignment with team and individual learning objectives.

MAGIC answers this need by helping training developers find, organize, and curate resources aligned with desired learning objectives. MAGIC analyzes source documents and extracts excerpts of content that aligns with specified learning objectives, and performs this alignment for both individual and team learning objectives. Moreover, MAGIC identifies content associated with specific roles within a team. A schematic depiction of MAGIC is shown in Figure 1.

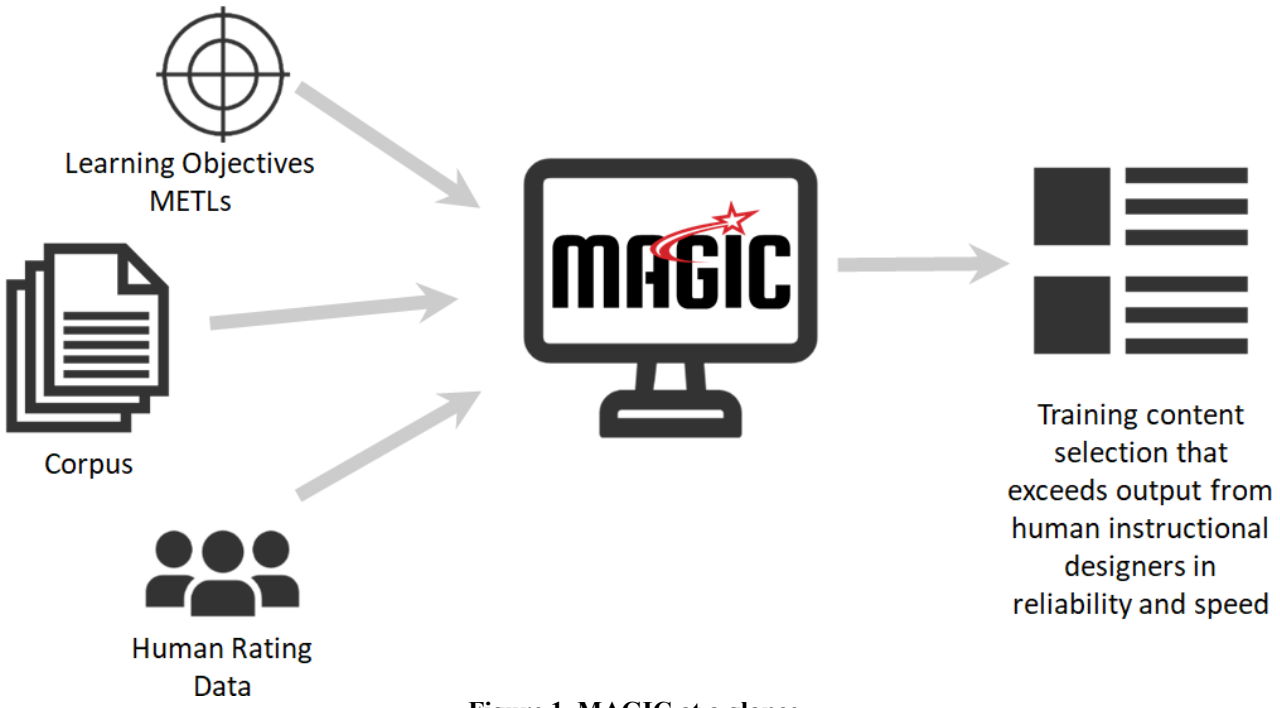


Figure 1. MAGIC at a glance.

STREAMLINING TRAINING DEVELOPMENT BY MATCHING CONTENT TO OBJECTIVES

MAGIC supports three tasks in a generalized GIFT authoring workflow (Figure 2).

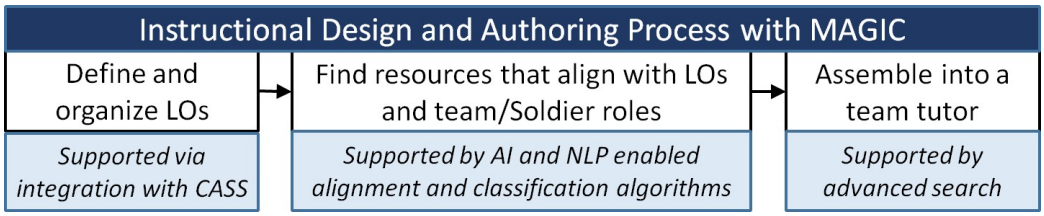


Figure 2. Primary authoring process tasks supported by MAGIC.

Using the MAGIC prototype UI, a training developer provides a list of learning objectives and selects the target library (or corpus of documents) to be analyzed as shown in Figure 3. For our initial demonstration of the MAGIC algorithms, we drew learning objectives from battle drills in the Maneuver domain; for the library we used the Central Army Registry (CAR) and the Milgaming portal's Training Support Packages (TSPs) to create a collection of over 1,200 documents.

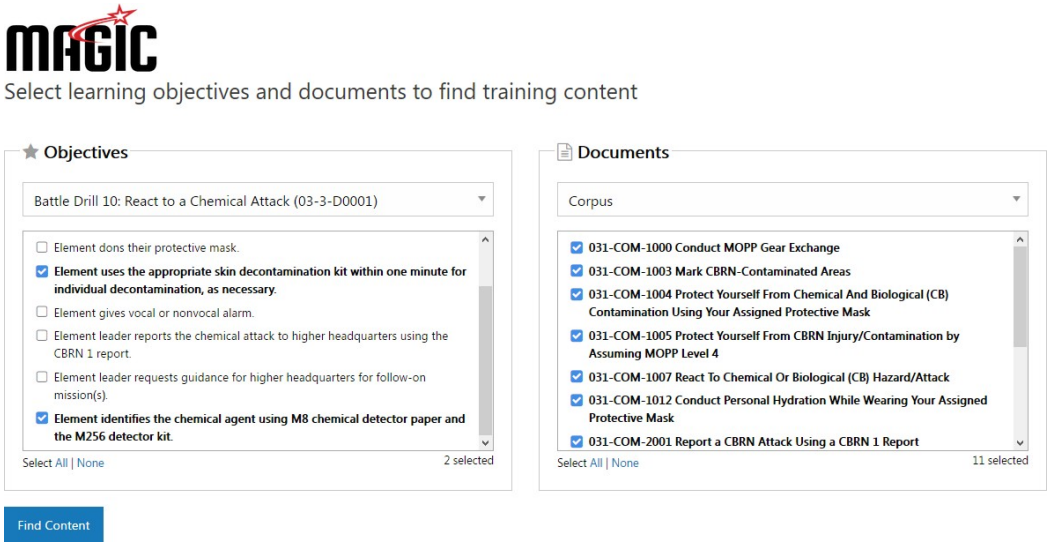


Figure 3: Selecting LOs and corpus documents to configure a content analysis.

MAGIC then generates a collection of text excerpts from across the selected documents, each tagged by the learning objectives, individual or team types, and team roles the excerpt aligns with. In the current demonstration interface, these results may be viewed, filtered, and compared with human rater results when available (Figure 4). In future work, the toolset will offer more flexible export packaging options designed to integrate into GIFT repository search and authoring components using the MAGIC API.

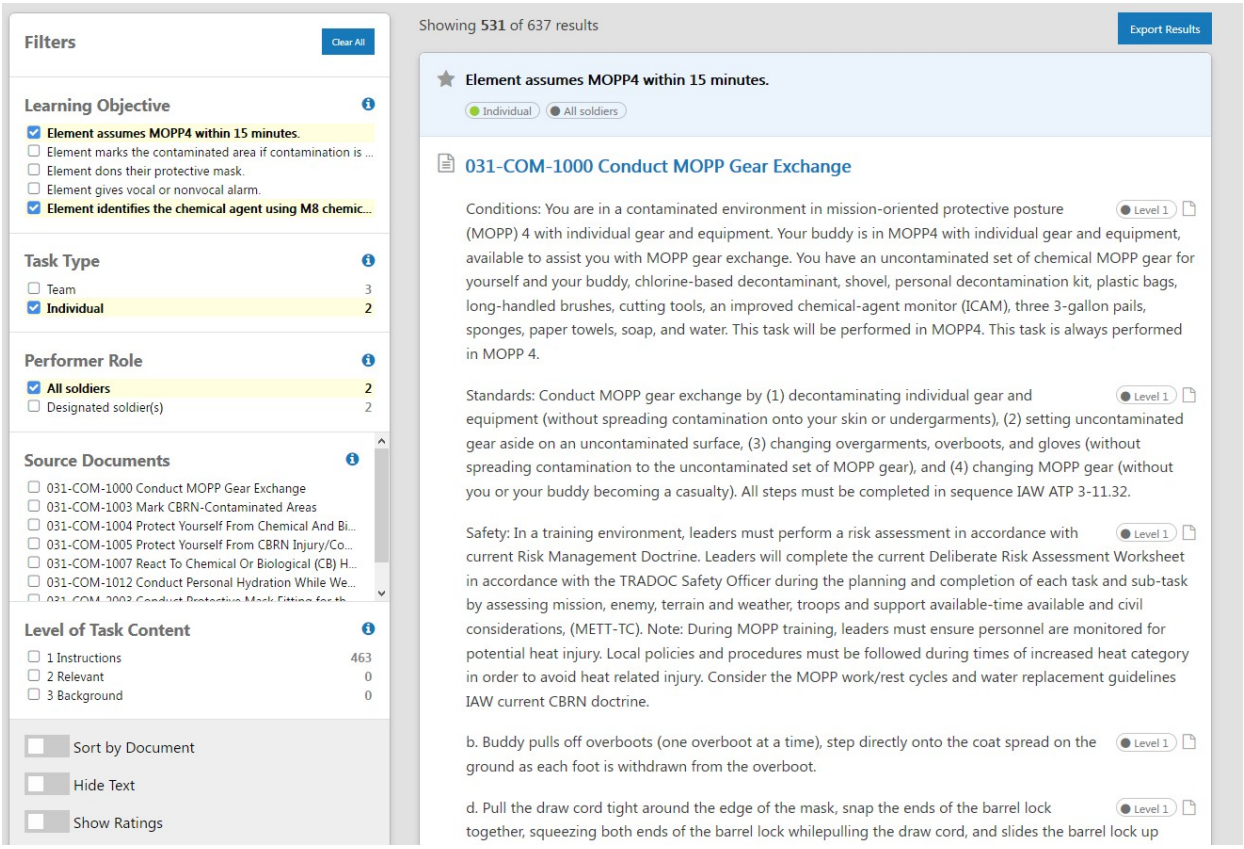


Figure 4. Filtering content by LO, task type, and role

MACHINE LEARNING: THE MAGIC BEHIND MAGIC

MAGIC uses ML and NLP techniques to train algorithms that associate content with learning objectives, tag content as having individual or team relevance, and associate content with specific team roles when applicable. We developed three sets of ML models for our initial research and testing: (1) *unsupervised general* models trained using Wikipedia and the New York Times Annotated Corpus to map concepts; (2) *unsupervised domain-specific* models trained with military-sourced documents to define domain-specific concepts; (3) *supervised, domain-specific* models trained with human-tagged data from a team of instructional designers and subject-matter experts to enhance outcomes.

In the case of the battle drill use cases, we manually created learning objectives (LOs) outlined as hierarchical task procedures, based on original document text, and manually tagged content with task type and role as depicted in Figure 5. The manually tagged LOs were used to train the ML algorithms for task type and role detection.

COMPETENCY ID	PARENT ID	Standard or Competency	Team or Individual	Role(s)
BD10-S1		*All Soldiers don their protective mask within nine seconds (or fifteen seconds for masks with a hood).*	*Individual*	*All soldiers*
BD10-S1-T1	BD10-S1	Element dons their protective mask.	Individual	All soldiers
BD10-S1-T2	BD10-S1	Element gives vocal or nonvocal alarm.	Team	Designated soldier(s)
BD10-S1-T3	BD10-S1	Element uses the appropriate skin decontamination kit within one minute for individual decontamination, as necessary.	Team	All soldiers
BD10-S2		*Soldiers assume MOPP4 within eight minutes.*	*Individual*	*All soldiers*

Figure 5. Example learning objectives for a battle drill.

To create the tagged data we used a team of three human raters with instructional design, research, and military backgrounds, led by an expert in instructional design. Raters were trained on the rating task, which included

scoring relevance of sections of content to a learning objective and tagging with individual/team and team role identifiers. The resulting tagged data set consists of 3,132 tagged items and was segmented into two corpora: one for training the supervised learning models, and one for evaluating performance of all three ML model sets. The average interrater reliability (n=3) was 81.6% for text selection and extraction, 87.8% for distinguishing team and individual content, and 78% for identifying team roles.

NOVEL SOLUTIONS

A challenge MAGIC addresses is matching content excerpts to a learning objective (typically a short text string) rather than to a topic (typically supported by larger amounts of descriptive text). To address this difficulty, we extended existing work in word embedding approaches (e.g. Word2Vec, GLoVe) (Mikolov et al, 2013; Pennington et al, 2014) , to develop a new technique we refer to as *concept embedding*. The approach first involves parsing an input corpus of documents to detect entities and relations as short phrases (rather than as individual words) using TensorFlow- or SyntaxNet-style dependency parsing along with traditional ontological approaches (Goldberg & Levy, 2014). In the next step, we build corpus models using the resulting dependency trees as the input into distinct entity and relation embedding models, where 'concepts' are defined as tight clusters of phrases in the resulting vector spaces (Levy & Goldberg, 2014). By mapping entities and relations separately, and then linking them through a combined (modified W2V-SG) model, we are able to instantiate concepts as tight clusters of phrases that exist in the resulting entity and relation vector spaces. For example, this approach might instantiate the concept "*Santa Claus*" as associated with "*Jolly Old St. Nick*" and "*the fat man in the red suit*." (Li et al, 2016, Shalaby et al, 2018),

This concept embedding approach gives MAGIC the ability to extract a richer description of meaning from very short text strings (namely, learning objectives). In our use case, the approach is applied in multiple steps to perform excerpt extraction:

- Extract entities and relations from the LOs
- Generate an *embedding space*
- Map entities to concepts
- Use any available context to disambiguate between concepts
- Map documents to the concept space (both concept and topic levels)
- Match concepts in each LO to concepts in the corpus
- Rank results based on match to both entity-concepts & relation-concepts of given LO

In order to discriminate between individual or team LO types, we applied a hybrid ML approach that was combined with syntactic-semantic patterns (Kelsey et al, 2017). On the ML side, we first extracted the semantic and syntactic features and tested using Naïve Bayes and Support Vector Machine (SVM) classification techniques which produced similar results. However, these two approaches were more accurate and required less training data than either a Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) implementation. On the Syntactic-Semantic side, we extracted combined syntactic-semantic features using SyntaxNet with TensorFlow, and then matched using the pattern library. We achieved the best results by applying both the ML and Syntactic-Semantic Pattern approaches and then using context-specific heuristics (where 'context' is derived from features of the source document and larger source text) to resolve any disagreements when selecting the team or individual label.

When identifying an appropriate team role for an excerpt, we determined that the link to LOs/competency frameworks can provide important role implications as well as provide a predefined list of possible roles. Our approach was to expand each role into a Concept using the Concept Embedding Model, and then to apply a similar matching approach. We continue to take steps to improve results with role assignment by using human-labelled

data to detect discourse and semantic-syntactic markers for a list of common domain-specific roles. The application of a supervised learning layer using human-tagged samples is expected to further enhance MAGIC outcomes, with a goal of achieving accurate extractions and tag selections more often than the human raters.

PRELIMINARY RESULTS

To provide early metrics of MAGIC's performance, we used the second set of labeled data as a test set. Both the training and test sets comprised approximately 5,000 comparisons of a text excerpt to a learning objective, and each task was completed by the three independent raters. Interrater reliability was 81.6%.

	Unsupervised Domain-General Model	Unsupervised Domain-General Model + Unsupervised Domain-Specific Model	Unsupervised Domain-General Model + Unsupervised Domain-Specific Model + Supervised Domain-Specific Model
Match 1: Machine agrees with raters when all humans agree	0.827	0.923	0.986
Match 2: Machine agrees with the majority when humans disagree	0.514	0.723	0.769
Match Total	0.762	0.871	0.948

Figure 6. Preliminary results for each of MAGIC's ML models.

The results (Figure 6) demonstrate the algorithms performing slightly below human performance when using only the domain-general unsupervised model, at or near human performance when adding the unsupervised domain-specific model, and slightly above human performance when adding the supervised domain-specific model.

CONCLUSIONS AND FUTURE WORK

With preliminary results already meeting human-rater levels of reliability using the combined unsupervised general and domain specific models, and with the addition of a supervised domain-specific model performing better than the human raters, the MAGIC approach is showing promising results and a path for continued enhancement. Based on these early findings, we see the potential for automated content discovery using LO auto-alignment and text extraction will result in faster, scalable team training development processes. Integration of MAGIC services into the GIFT authoring workflows will propel reuse of training materials, while helping training developers overcome the challenges of distinguishing content supporting team or individual learning and aligning content with specific team roles.

Our next steps in the MAGIC project will include creating a supervised domain-specific model for assigning team roles; incorporating non-text content (such as metadata or automated transcriptions); designing a MAGIC services API; testing and evaluation of MAGIC with authors of team training simulations; and the integration of MAGIC services with Army-selected authoring/CMS/LMS tools.

REFERENCES

Bonner, Desmond; Gilbert, Stephen B.; Dorneich, Michael C.; Winer, Eliot; Sinatra, Anne M.; Slavina, Anna; MacAllister, Anastacia; and Holub, Joseph, (2016) "The Challenges of Building Intelligent Tutoring Sys- tems for Teams". *Industrial and Manufacturing Systems Engineering Conference Proceedings and Posters*. 23. Retrieved from http://lib.dr.iastate.edu/imse_conf/23

Fletcher, J. D., & Sottilare, R. A. (2017). Shared mental models in support of adaptive instruction of collective tasks using GIFT. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593- 017-0147-y.

Gilbert, S., Slavina, A., Dorneich, M., Sinatra, A., Bonner, D., Johnston, J., Holub, J., MacAllister, A., and Winer, E. (2017). Creating a team tutor using GIFT. *International Journal of Artificial Intelligence in Educa- tion.* https://doi.org/10.1007/s40593-017-0151-2.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word- embedding method. arXiv preprint arXiv:1402.3722.

Kelsey, E. Goetschalckx, R. Robson, E., Ray, F. & Robson, R. Automated Tools for Question Generation. United States Patent Application 20180260472

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (Vol. 2, pp. 302-308).

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016, July). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Develop- ment in Information Retrieval* (pp. 165-174). ACM.

McCormack, R., Kilcullen, T., Sinatra, A.M., Brown, T., & Beaubian, J. (2018). Scenarios for Training Teamwork Skills in Virtual Environments with GIFT. In Proceedings of the 6th Annual *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym6)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Salas, E. , Shuffler, M. L., Thayer, A. L., Bedwell, W. L. and Lazzara, E. H. (2015), Understanding and Improving Teamwork in Organizations: A Scientifically Based Practical Guide. *Human Resource Management*, 54: 599-622. doi:10.1002/hrm.21628

Shalaby, W., Zadrozny, W., & Jin, H. (2018). Beyond word embeddings: learning entity and concept representations from large scale knowledge bases. *Information Retrieval Journal*, 1-18.

Sinatra, A.M. (2018). Team Models in the Generalized Intelligent Framework for Tutoring: 2018 Update. In Proceedings of the 6th Annual *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym6)*.

Sottilare, R., Holden, H., Brawner, K., & Goldberg, B. (2011). Challenges and emerging concepts in the develop- ment of adaptive, computer-based tutoring systems for team training. In *Proceedings of the Interserv- ice/Industry Training Systems & Education Conference*, Orlando, December 2011.

Sottilare, R., Holden, H., Goldberg, B., & Brawner, K. (2013). The Generalized Intelligent Framework for Tutoring (GIFT). In Best, C., Galanis, G., Kerry, J. and Sottilare, R. (Eds.) *Fundamental Issues in Defence Simula- tion & Training*. Ashgate Publishing.

Sottilare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017a). An updated concept for a generalized intelligent framework for tutoring (GIFT). Orlando: US Army research laboratory. May 2017. https://doi.org/10.13140/RG.2.2.12941.54244.

Sottilare, R. A., Burke, C. S., Salas, E., Sinatra, A. M., Johnston, J. H., & Gilbert, S. B. (2017b). Towards a design process for adaptive instruction of Teams: A Meta-Analysis. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-017-0146-z.

Sottilare, R., Graesser, A., Hu, X., and Sinatra, A. (Eds.). (2018). *Design Recommendations for Intelligent Tutoring Systems: Volume 6 - Team Tutoring*. Orlando, FL: U.S. Army Research Laboratory. ISBN 978-0-9977257- 4-2. Available at: https://gifttutoring.org/documents/ and on Google Play.

ABOUT THE AUTHORS

Dr. Benjamin Bell *is the Principal Investigator on the MAGIC project and president of Eduworks. He has been leading funded research in education, training and simulation for over twenty years, with an emphasis on defense and national security applications. He holds a Ph.D. from Northwestern University in Artificial Intelligence, Master's degrees from Embry Riddle and Drexel University, and a Bachelor's degree from the University of Pennsylvania.*

Dr. Keith Brawner *is a senior researcher for the U. S. Army Combat Capability Development Command Soldier Center at the Simulation and Training Technology Center (CCDC-SC-STTC), and is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT). He has 13 years of experience within U.S. Army and Navy acquisition, development, and research agencies. He holds a Masters and PhD degree in Computer Engineering with a focus on Intelligent Systems and Machine Learning from the University of Central Florida. His current efforts are on artificial intelligence for the Synthetic Training Environment Simulation and Network Compression. He manages research in adaptive training, semi/fully automated user tools for adaptive training content, and architectural programs towards next-generation training.*

Elliot Robson *leads research and development for Eduworks corporation as General Manager. He has fifteen years of experience in training technology and has served as PI on projects using GIFT.*

Debbie Brown *acts as Learning Engineer, Instructional Designer, and Software Developer at Eduworks, and has been active in eLearning research and development for 20 years in academic, workforce, and government/military applications. Since joining Eduworks, she has served on multiple projects using GIFT.*

Elaine Kelsey *is a research engineer at Eduworks, focusing on applications of natural language processing and machine learning in topic and concept detection and automated text generation. She designed and led the development of Eduworks' automated question generation algorithm, and is currently developing solutions in human- interpretable machine learning. She has multiple Bachelors and Masters degrees in computer science, linguistics, and molecular biology, and is currently working towards a PhD in Computational Linguistics.*

Modeling and Visualizing Team Performance using Epistemic Net- work Analysis

Zachari Swiecki[1], A. R. Ruis[1], David Williamson Shaffer[1,2]
University of Wisconsin-Madison[1], Aalborg University Copenhagen[2]

INTRODUCTION

A key goal of The Generalized Intelligent Framework for Tutoring (GIFT) project is to scaffold the development of tutoring scenarios that support team-training and assessment (Sinatra, 2018). As Ruis and colleagues (Ruis, Hampton, Goldberg, & Shaffer, 2018) argue, a critical component of team training and assessment is the ability to *model* team performance. In this paper, we describe a specific approach to modeling team performance at both the individual and team levels that prior work suggests is more valid than extant approaches. Moreover, we argue that modeling team performance is not sufficient on its own—given the goals of team tutors, we also need visualizations that effectively summarize team performance and provide actionable information at both the team level and the individual level. Here, we describe the design of a team-tutoring *dashboard* that would allow tutors to monitor individual and team performance in real-time. This system could inform assessment or guide the delivery of feedback either in real-time or after the tutoring scenario is complete.

TEAM PERFORMANCE

When individuals on teams solve problems, their processes include (a) actions toward accomplishing a task and (b) actions toward managing the processes of collaboration. Thus, team processes are not simply the sum of individual actions; rather, individual actions *interact* with one another, creating a context independent of any single individual. As interactions unfold, they contribute to the *common ground*, or the shared knowledge and experience that exists between people when they interact (Clark, 1996). As a result, the discourse of the team is *interdependent*: the actions of one individual impact the actions of others on the team. Moreover, team processes have an important *temporal* dimension: team processes unfold in time and are interpreted with respect to the immediately preceding actions—or *recent temporal context*—not the entire history of team interaction (Reimann, 2009; Suthers & Desiato, 2012).

This complexity suggests that valid models of team performance at the *team* level should account for relationships among the recent contributions of the team, and valid models at the *individual* level should account for relationships between a given individual's contributions and the recent contributions of the rest of the team.

Modeling Team Performance

Despite these suggestions, many extant modeling approaches still employ *coding-and-counting* (Chi, 1997; Suthers, 2006). At the team level, this involves aggregating behavioral markers or codes over the entire history of a team task or scenario, ignoring temporal aspects of team processes. Similarly, coding- and-counting at the individual level ignores temporality, and because it separates the processes of individuals from the processes of the team, it also ignores the interactive and interdependent aspects of individual contributions (Csanadi, Eagan, Shaffer, Kollar, & Fischer, 2018). More nuanced analyses are often conducted with techniques that model frequent sequential events in data, such as *sequential pattern mining*; however, similar to coding-and-counting, such techniques can only model individuals irrespective of the team (Swiecki, Lian, Ruis, & Shaffer, in press [A]).

An alternative approach that can account for these critical aspects of team processes at the team and individual levels is *epistemic network analysis* (ENA) (Shaffer, 2017; Shaffer, Collier, & Ruis, 2016; Shaffer & Ruis, 2017). Specifically, ENA models team activity by identifying categories of action, communication, cognition, and other relevant features and characterizing them with appropriate coding schemes into smaller sets of domain-relevant nodes. The weights of the connections among network nodes (i.e., the association structure of key elements in the domain) are then computed and visualized. Critically, ENA models team actions and interactions in such a way that it is possible to *extract information about each team member's contributions to team performance*.

ENA uses statistical and visualization techniques to enable comparison of the salient properties of different networks, including networks generated by different teams or by teams at different points in time, teams in different spatial locations, or teams engaged in different activities. These salient properties are modeled not just in terms of the general structure of the networks, but ENA also extracts properties relevant to the actual content of the network.

In other words, ENA can analyze (a) what teams are doing, (b) how they are thinking, (c) what role individuals are playing in team performance, and (d) how teams compare to one another in the context of real problem solving. Moreover, prior work using ENA to model the performance of U.S. military teams—a key domain of interest for GIFT—has shown that ENA has both statistical and interpretive advantages compared to coding-and-counting and sequential pattern mining (Swiecki et al., in press [A]; Swiecki, Ruis, Farrell, & Shaffer, in press [B]).

TEAM-TUTORING DASHBOARD

While models such as those produced by ENA are useful tools for examining team performance, they are designed primarily for researchers. As such, their affordances are not necessarily aligned with the goals of other audiences that have an interest in understanding team performance, such as tutors or the teams themselves (Swiecki & Shaffer, 2018). For example, an important goal of researchers is to advance their understanding of phenomena or make predictions, and they are trained to understand and use complex models and visualizations to help them do so. Tutors, on the other hand, need to assess performance and guide interventions, and they may lack the training required to effectively use complex models and visualizations. In turn, tutors need tools that quickly highlight the teams or individuals that need their attention the most, while also providing them information that can guide their interventions.

As a first step toward integrating such a system with GIFT, we have created preliminary designs of a team-tutoring dashboard. This dashboard uses simplified ENA models to provide actionable information on the performance of teams and individuals. These designs are based on prior work in which we successfully designed, built, and implemented an ENA-driven team performance dashboard in a simulation-based learning environment (Herder et al., 2018). The ENA models presented in the designs are based on prior work by Swiecki and colleagues (in press [A], in press [B]). We summarize the data and relevant models from this work in more detail below.

ENA Models of Team Performance

As part of the Tactical Decision Making Under Stress project, sixteen teams participated in training scenarios to test the impact of a new decision-support system on team performance in the context of air defense warfare (Johnston, Poirier, & Smith-Jentsch, 1998). During the scenarios, teams needed to detect and identify ships and aircraft (referred to as *tracks)*, assess whether they were threats, and decide how to respond. Each team consisted of six members who held either a leadership role, such as the Commanding Officer (CO), or a support role, such as the Electronic Warfare Supervisor (EWS). The dataset consists of transcripts of team communications and performance scores for each team.

To create the ENA models, we developed and validated an automated coding scheme that captured the critical aspects of the team task. After coding, we used ENA to create models at the team and individual level. The ENA algorithm uses a sliding window to construct a network for each turn of talk in the data, showing how codes in the current turn of talk are connected to codes within the recent temporal context. In other words, ENA defines a connection between codes as their co-occurrence within a specific number of turns of talk. To create networks for each unit of analysis, ENA aggregates the networks associated with their turns of talk. In this way, ENA can model the network of connections that each team or individual makes between concepts and actions while taking into account the recent actions of others (Siebert-Evenstone et al., 2017).

Two coordinated representations are produced for each team or individual network: an ENA score and a weighted network graph. ENA uses a dimensional reduction via spectral value decomposition (SVD) to create an ENA score for each team or individual that summarizes their network of connections. These scores give their location in the *ENA space* created by the dimensional reduction. Typically, this dimensional reduction maximizes the variance accounted for by each dimension. However, ENA can also combine SVD with a hyperplane projection such that

the first dimension maximizes the variance between the means of two subpopulations—for example, high and low performing teams—present in the data.

The nodes of the weighted network graphs correspond to codes, and the edges are proportional to the relative frequency of connection between two codes. The positions of the network graph nodes are fixed across networks, and their positions are determined by an optimization algorithm that minimizes the difference between the ENA scores and their corresponding network centroids. This relationship implies that ENA scores toward the extremes of a dimension have network graphs with strong connections between nodes located on the extremes. As a result, dimensions in this ENA space distinguish teams in terms of connections between codes whose nodes are located at the extremes. In addition, ENA can produce network difference graphs which subtract the edge weights of two networks to show the connections that are strongest in one network relative to another.

At the team level, our analysis suggested that high performing teams made frequent connections between *tactical information*, such as track behavior and track detection, and *tactical actions* such as combat orders. Low performing teams made relatively frequent connections to seeking information, suggesting that they had difficulty maintaining situational awareness (Figure 1, left).

At the individual level, our analysis suggested connections for leadership and support roles that distinguished those on high and low performing teams. Connections that distinguished individuals in leadership roles were very similar to those that distinguished high from low performing teams, so we do not describe them in detail here. Individuals in support roles on high performing teams made frequent connections to status updates, suggesting that they played a critical role in updating the team on the evolving tactical situation. Individuals in support roles on low performing teams made more frequent connections to seeking information which suggests that they were focused on repairing the team's understanding of the tactical situation (Figure 1, right).

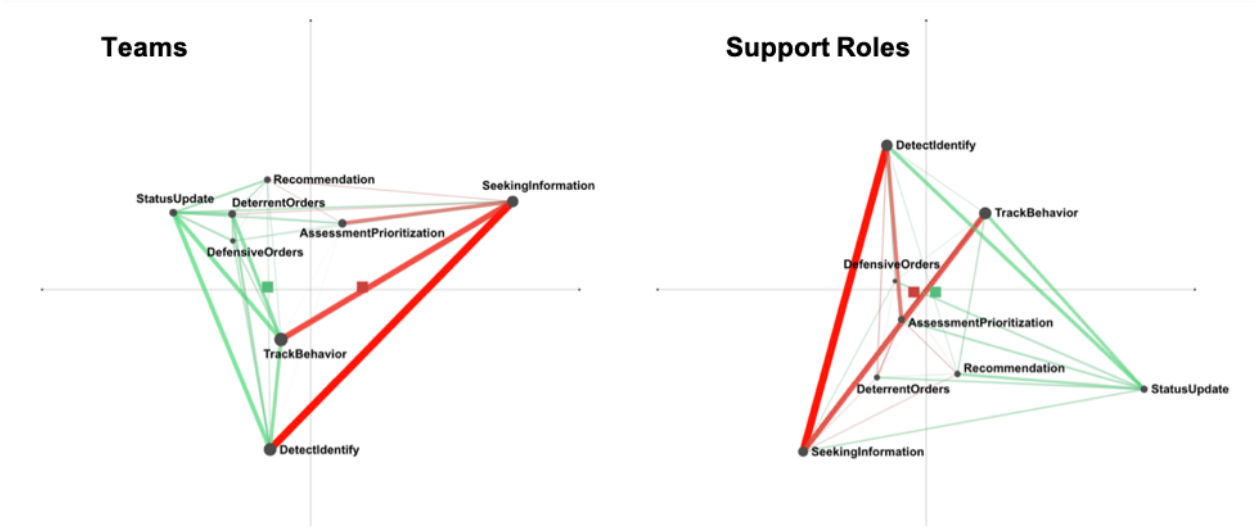


Figure 1. ENA Models of Team Performance: ENA difference network between high performing teams (left, green) and low performing teams (left, red). ENA difference network between individuals in support roles on high performing teams (right, green) and individuals in support roles on low performing teams (right, red). Difference network for individuals in leadership roles not shown due to similarity with team network.

In the next section, we describe our process for integrating ENA models in to a team-tutoring dashboard using these data and results as an example.

Dashboard Designs

The proposed dashboard features a performance *overview* of all teams and individuals and the ability to *drill down* to more specific information about the performance of a team or individual. In the performance overview (Figure 2), each row represents a different individual in a particular role (e.g., CO, EWS, etc.) grouped by their team; each column represents a different training scenario. For a given scenario, high performing individuals are indicated by

a green circle, average performers are shown in yellow, and low performers in red. Team members with no activity a represented by an empty circle, and those with no relevant activity by a grey circle. High, average, and low performance indicators are determined by thresholds on the distribution of ENA scores from either the leadership or support ENA model. For these designs, thresholds were set at the first and third quartile of the distributions, but in the general case, they would be customizable.

This overview has several affordances for team tutors. First, it provides a quick reference for how teams or individuals are performing, and thus directs attention to those who may need an intervention. Second, it presents the performance of a team or individual in the context of others, which facilitates comparisons.

Third, the horizontal axis allows tutors to track performance over time to examine and compare trends. Finally, the vertical axis allows tutors to track performance across a given scenario to examine whether that scenario is more or less difficult than others. Such affordances are important because they scaffold decisions about whether an intervention is necessary and what kind of intervention to provide. For example, a CO who has high performance across all scenarios but one would likely need a different intervention than a CO whose performance was more variable over time.

Figure 2. Performance Overview

By clicking a team or individual, tutors can drill down to see more detailed information about their performance. For example, Figure 3 shows a simplified *network model* of one team's performance at the end of a training scenario. Green connections between codes are characteristic of high performing teams; red connections are characteristic of low-performing teams. In other words, high-performing teams will have a higher frequency of green connections relative to red.

To create this simplified network model, we selected the connections from the ENA models described above that explained the most variance between high and low performing teams in the dataset—that is, the connections at the extremes of the first dimension in the ENA space. Unlike the models described above, the node placement of the simplified networks is designed for easy comprehension, with nodes connected by green (i.e., indicative of high performance) connections placed at the top of the display. In addition, this drill down view shows the team *activity* represented in the network model—in this case, the coded team transcript. Turns of talk in the activity record with black circles have a code present in the turn. Below the network is a *description* of the connections present in the network (See Figure 5 for more details). Note that placeholder text is used for the activity record and network description in all designs except those in Figure 5.

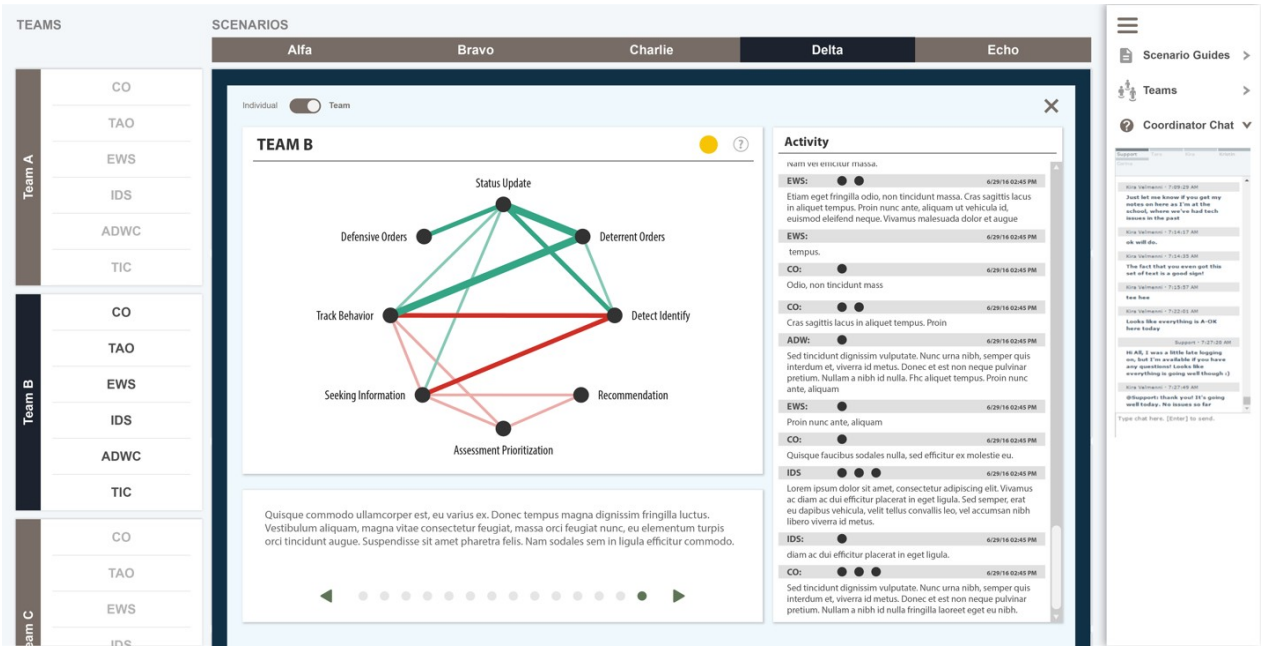


Figure 3. Team Network Model

Tutors can access similar drill downs for each individual on the team. For example, Figure 4 shows the network visualization of this team's EWS, who holds a support role, at the end of a training scenario. Green connections are characteristic of high-performing individuals in support roles; red connections are characteristic of low performing individuals in support roles. To create simplified network visualizations for individuals in either leadership or support roles, we selected the connections from the ENA models described above that explained the most variance between individuals in those roles who were on high performing versus low performing teams. Node placements match the positions of the team networks to maintain visual consistency between individuals and teams.

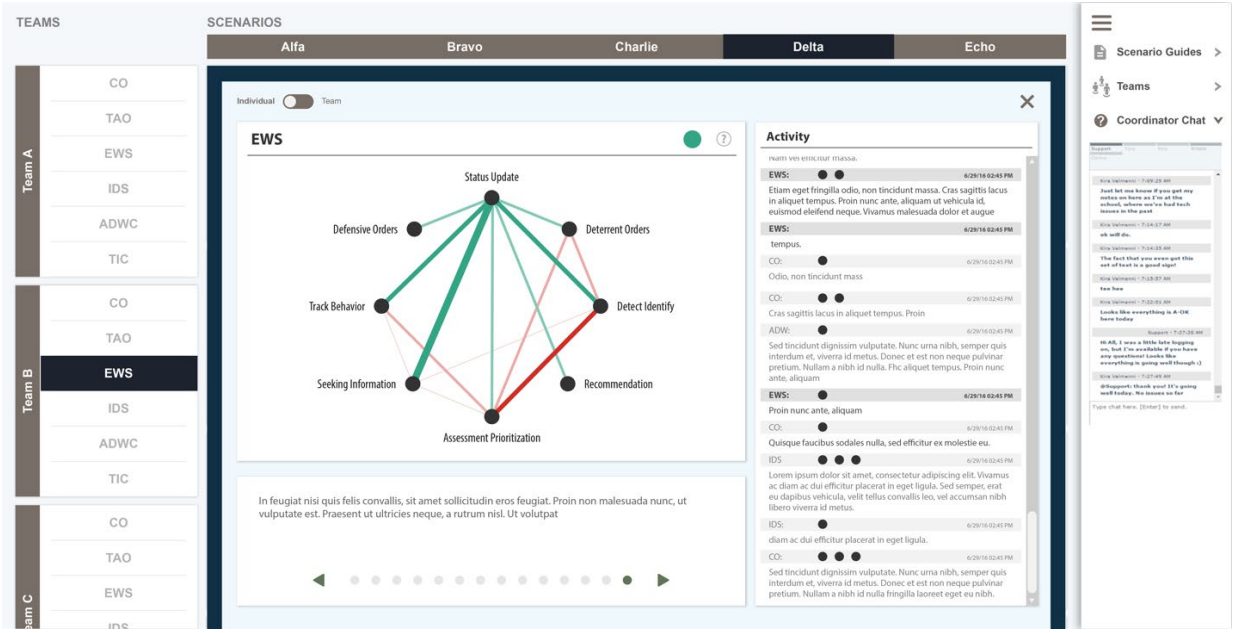


Figure 4. Network Model

Tutors can use the arrows in the description below the network model to step through the scenario in time and review each connection (and the activity contributing to the connection) made by the individual or team. For example, in Figure 5, we can see the first connection made by the team's CO in this training scenario. Here, the CO is responding to tactical information from the Tactical Action Officer (TAO) with an order.

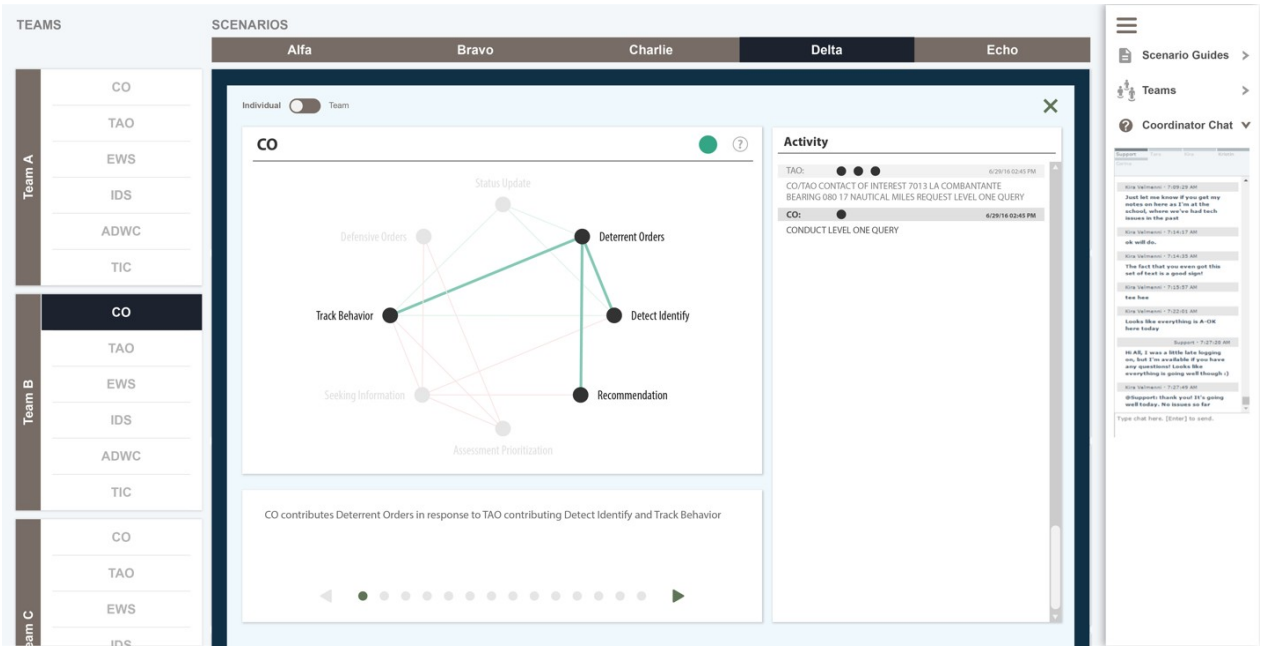


Figure 5. Network Review

In addition to stepping through the network model, tutors can examine the activity contributing to a connection by clicking the connection in the network model. As shown in Figure 6, clicking a connection in the network highlights the most recent activity in which the connection occurred.

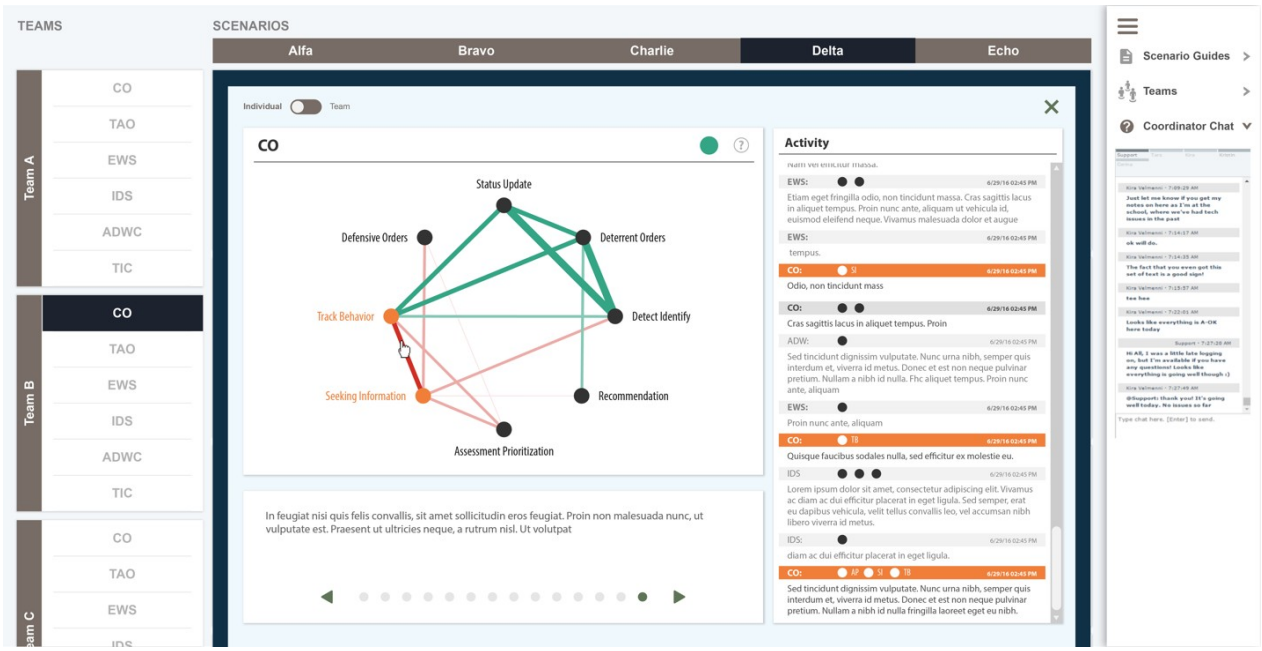


Figure 6. Connection Inspection

Similarly, as shown in Figure 7, tutors can also click a segment of activity in the activity record to highlight any connections that may have occurred within the recent temporal context of that segment.

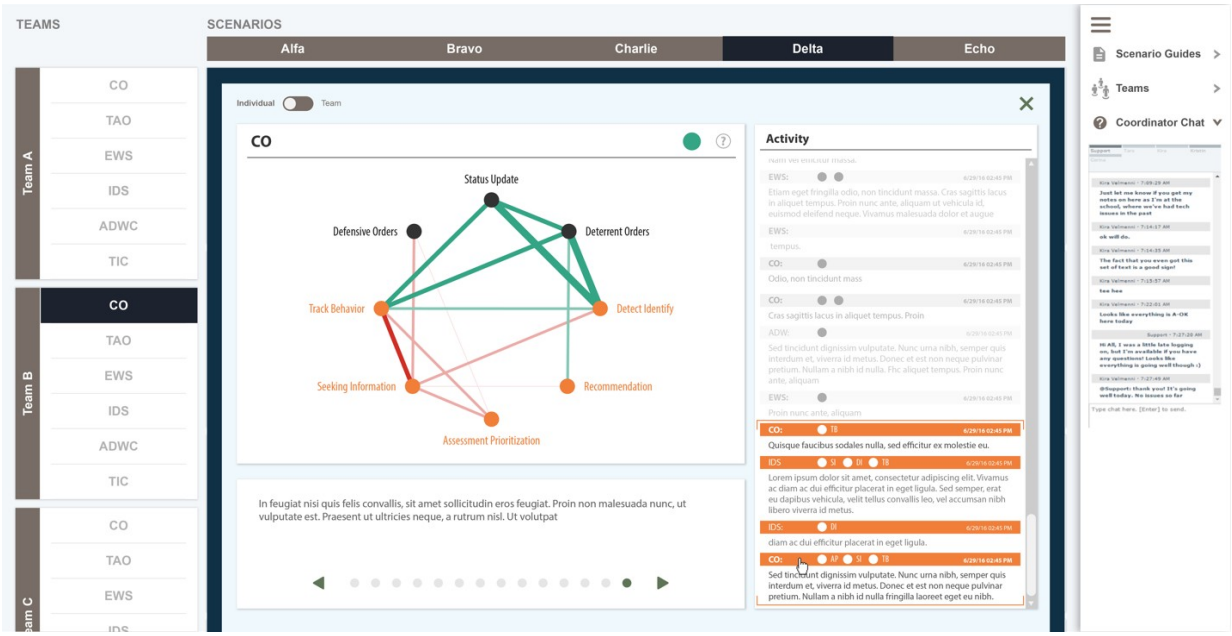


Figure 7. Activity Inspection

Features such as stepping through the network model and investigating connections and their corresponding activity can facilitate after-action reviews by the tutor with teams or specific individuals.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

The work presented here suggests two recommendations for GIFT. First, team-tutoring assessments in GIFT should include models, such as ENA, that account for the interactive, interdependent, and temporal nature of team processes at both the team and individual levels. Second, in order to be successful, such assessments should include actionable visualizations that help tutors monitor, assess, and provide feedback on team performance. The designs described above are one proposal for such an approach.

While these designs used specific data from a particular domain for illustrative purposes, the approach is agnostic to both the kind of data collected and the domain from which it comes. The only constraints are that the data consists of a machine readable record of ordered events, which may be talk, gestures, mouse- clicks, or other actions, and that there exists a reliable automated coding scheme for the data. In cases where the data is not in a textual format, such as raw audio or video recordings, the system would need a means of converting the data in real-time into a format which could be automatically coded. Moreover, while the network models described here were data driven, it is also possible for authors to specify a priori the connections that distinguish high and low performance for teams and individuals.

Our future work will include adapting the dashboard designs to different team structures, such as squads or platoons, and mapping the components of these designs to the inputs of *domain knowledge files* that manage assessment and pedagogical requests for teams or individuals during a GIFT-managed scenario.

ACKNOWLEDGEMENTS

This work was funded in part by the National Science Foundation (DRL-1661036, DRL-1713110), the U.S. Army Research Laboratory (W911NF-18-2-0039), the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin- Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

REFERENCES

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, *6*(3), 271–315.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Csanadi, A., Eagan, B., Shaffer, D. W., Kollar, I., & Fischer, F. (in press). When coding-and-counting is not enough: Using Epistemic Network Analysis (ENA) to analyze verbal data in CSCL research. *International Journal of Computer-Supported Collaborative Learning*.

Herder, T., Swiecki, Z., Fougt, S. S., Tamborg, A. L., Allsopp, B. B., Shaffer, D. W., & Misfeldt, M. (2018). Supporting teacher's intervention in student's virtual collaboration using a network-based model. *Proceed- ings of the International Conference on Learning Analytics*, 21–25. Sydney, Australia.

Johnston, J. H., Poirier, J., & Smith-Jentsch, K. A. (1998). Decision making under stress: Creating a research methodology. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 39–59). Washington, D.C.: American Psychological Association.

Reimann, P. (2009). Time is precious: Variable- and event-centered approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, *4*(3), 239–257. https://doi.org/10.1007/s11412-009-9070-z

Ruis, A. R., Hampton, A. J., Goldberg, B. S., & Shaffer, D. W. (2018). Modeling processes of enculturation in team training. In R. Sottilare, A. Graesser, & A. M. Sinatra (Eds.), *Design Recommendations for Intelligent Tu- toring System: Volume 6 - Team Tutoring* (pp. 45–51). Orlando, FL: U.S. Army Research Laboratory.

Shaffer, D. W. (2017). *Quantitative ethnography*. Madison, WI: Cathcart Press.

Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, *3*(3), 9–45.

Shaffer, D. W., & Ruis, A. R. (2017). Epistemic network analysis: A worked example of theory-based learning analytics. In C. Lang, G. Siemens, A. F. Wise, & D. Gasevic (Eds.), *Handbook of learning analytics* (pp. 175–187). Society for Learning Analytics Research.

Siebert-Evenstone, A., Arastoopour Irgens, G., Collier, W., Swiecki, Z., Ruis, A. R., & Williamson Shaffer, D. (2017). In Search of Conversational Grain Size: Modelling Semantic Structure Using Moving Stanza Win- dows. *Journal of Learning Analytics*, *4*(3), 123–139. https://doi.org/10.18608/jla.2017.43.7

Sinatra, A. M. (2018). Team Models in the Generalized Intelligent Framework for Tutoring: 2018 Update. In R. Sottialre (Ed.), *Proceedings of the Sixth Annual GIFT Users Symposium* (pp. 157–161). Orlando, FL.

Suthers, D. D. (2006). Technology affordances for intersubjective meaning making: A research agenda for CSCL. *International Journal of Computer-Supported Collaborative Learning; New York*, *1*(3), 315–337. http://dx.doi.org.ezproxy.library.wisc.edu/10.1007/s11412-006-9660-y

Suthers, D. D., & Desiato, C. (2012). Exposing chat features through analysis of uptake between contributions. *System Science (HICSS), 2012 45th Hawaii International Conference On*, 3368–3377. IEEE.

Swiecki, Z., Lian, Z., Ruis, A. R., & Shaffer, D. W. (in press) [A]. Does order matter? Investigating sequential and cotemporal models of collaboration. *Proceedings of the 13th International Conference of Computer- Supported Collaborative Learning*. Presented at the Lyon, France. Lyon, France.

Swiecki, Z., Ruis, A. R., Farrell, C., & Shaffer, D. W. (in press) [B]. Assessing individual contributions to collabora- tive problem solving: A network analysis approach. *Computers in Human Behavior*.

Swiecki, Z., & Shaffer, D. W. (2018). Toward a taxonomy of team performance visualization tools. In J. Kay & R. Luckin (Eds.), *Rethinking Learning in the Digital Age: Making the Learning Sciences Count*: *Vol. III* (pp. 144–151).

ABOUT THE AUTHORS

Zachari Swiecki *is a Learning Sciences Ph.D. candidate at the University of Wisconsin–Madison. His work focuses on modeling and visualizing collaborative problem solving.*

Andrew R. Ruis *is a researcher at the University of Wisconsin–Madison. His work focuses on modeling deliberation and decision making in healthcare and educational contexts.*

David Williamson Shaffer *is the Vilas Distinguished Achievement Professor of Learning Science at the University of Wisconsin–Madison. His work focuses learning analytics, with an emphasis on models of collaborative problem solving.*

Integrating Gift, Competencies, Virtual Reality, And Biometrics To Present Training Perspectives On Gauging Current Squad Capability

Zach Heylmun[(1)], Mike Kalaf[(1)], Chris Meyer[(1)], Christofer Padilla[(2)], Lucy Woodman[(1)]
Synaptic Sparks, Inc. [(1)], Dignitas Technologies [(2)]

INTRODUCTION

ITSs have continued their conceptual evolution and implementation in parallel to modern technologies that can be configured to add value to ITSs' overall effectiveness. Along with the evolution of technology and software, the Department of Defense (DoD) and the U.S. Army have evolved their programmatic practices to include modernization priority updates for 2018-2019. As part of this volunteer effort; it is to experiment-with and add-value-to one of those modernization priorities, namely Soldier Lethality that the authors proposed that GIFT could be used to supplement.

GIFT was configured to integrate with and utilize Commercial-Off-The-Shelf (COTS) hardware for both VR solutions and external sensors in this effort to create an experimental scenario for modernized adaptive training. GIFT also had some of its internal user interfaces updated to display dynamic mastery and competency information using a CASS test database at https://cassproject.github.io/cass-editor/. By combining GIFT's capability to adapt training content with a competency standards system such as CASS, the authors hoped to enable formal experiments that measure training value as a system when compared to the individual software components alone.

Primarily, this effort created the framework with which to further experiment integrating GIFT with biometrics and virtual reality, along with other future IOT devices and software suites. The authors' teams consisted mainly of working professional engineers contributing to nonprofit efforts, but the direct team member roster did not contain Army Subject Matter Experts (SMEs) or behavioral scientists / doctors with which to form experimental hypothesis and validations. It was the authors' intent that the framework may now be tailored to suit specific scientific needs in the community having enabled the prototype functionality.

The framework was secondarily created to provide GIFT with another set of sensors and applications with which to integrate and perform future training scenarios with. The training scenario produced as part of the paper's effort was not constructed as an actual DoD course, but rather as an example on "the art of the possible" on how to use many of the technological evolutions that the IOT- style of hardware production has provided modern society. Sample IOT devices included with this experiment include pulse monitoring sensors, indicator lights, haptic motors, and real-time situational team knowledge simulation, all of which can be purchased through COTS providers for under $20 total per trainee.

Combining GIFT, VR, CASS, COTS IOT hardware, and the team's engineering experience, the authors created, to best effort, a breadth-first course containing a shallow dive into all of these different areas proving high levels of interoperability from GIFT. Proposed future experimentation variables include the type and fidelity of VR content, learner characteristics such as familiarity with VR and trainee career background, length of time in VR while in a scenario, difficulty of the scenario referencing the number of IOT devices selected, and the quality of results as different types of adaptation are used during a scenario. This project was not funded from any source and is released with full rights to any interested public entity. Even in this case where the quantity of training content was minimal, the results of the effort should prove to be of interest to the research community, the GIFT team, and possibly multiple branches of the military as training scenarios continue to modernize along the path of higher technology.

Parts List and Descriptions

This section describes the parts and components that were used in the making of the prototype system described in this paper. Interested community members are encouraged to request further information or specifications from any of the authors if desired.

GIFT Software Suite

At the time of this writing, GIFT 2019-1 has been released at www.gifttutoring.org/projects/gift/files. If the reader has not yet created a GIFT account to enable the download, registration is free by following the link to 'Register' on the web page. The GIFT 2019-1 download will allow the reader to install and configure their own local GIFT server for any purpose. Instructions for configuring a GIFT server and discussions on the matter can be found included with the download and on the Forum tab at the www.gifttutoring.org homepage.

CASS Online Test Database

By following the https://cassproject.github.io/cass- editor/ link, readers may explore the site to edit and configure their own competency and mastery framework. This server is maintained by the Advanced Distributed Learning (ADL) CASS team, but any user has full permissions to create and edit their own framework. Readers may search for 'GIFTSym7' to examine the framework created for this paper's effort. Readers may also register with the CASS project at https://www.cassproject.org, download the open source code from the referenced GitHub project, and build/configure/maintain their own CASS server.

Figure 1: Screenshot of CASS Framework for GIFTSym7 in CASS Editor Web Tool

MQTT Online Network Communication Server

A tool called Mosquitto (with two T's) is a lightweight communication protocol that follows a publish/subscribe model of network communications. Furthermore, a public test server is available at https://test.mosquitto.org. Readers may also download the server/client itself to setup their own instance on any available machine. The online server is free for the public to use, and was incredibly valuable as an online communication server through which to test messages being passed between the various components of the paper's system. While building the prototype system simply using 'localhost' methodologies was possible, the team wished to demonstrate the potential power and distributed nature of open API software such as GIFT, CASS, Unity VR, and modern IOT devices.

Unity

The Unity game engine was chosen to host the prototype virtual environment due to GIFT's pre-existing integrations and course content utilizing Unity. A personal version of Unity may be downloaded from https://unity.com for noncommercial users. It should be noted that GIFT is an agnostic ITS, meaning that there is

no restriction on which virtual environment that adaptive tutoring can be performed in. For instance, GIFT already contains a Course Object to manage many available interactions with Virtual Battle Space (VBS): https://bisimulations.com/products/virtual-battlespace. Proper configurations of network communications according to GIFT's open API instructions allow for any environment to become part of adaptive training and allow for virtual environment events to be monitored or injected according to GIFT course direction. New experimental features in GIFT that are currently being developed also allow for more generic communications and management of scenarios within virtual environments.

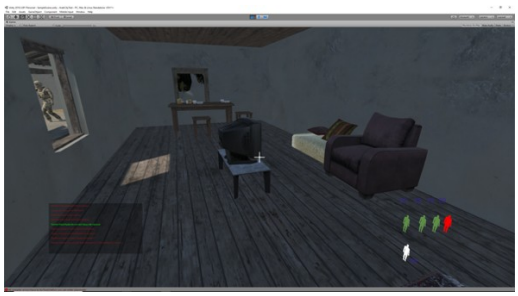


Figure 2: Unity Scenario Room with HUD

Learner Record Store – Learning Locker

Interfaces to a Learning Management System (LMS) and/or Learner Record Store (LRS) were not integrated as part of this paper's effort, but GIFT does contain existing interfaces to an LRS called Learning Locker, https://www.ht2labs.com/learning-locker/. When tracking permanent student performance, specifically as it relates to competency and mastery acquisition, GIFT can agnostically communicate with any open API LMS or LRS given software development time to create the simple translations of existing network messages. This capability will allow GIFT to store permanent individual and team performance measures and progressions accessible in the CASS database as Army Subject Matter Experts (SMEs) enter in competency information and relationships.

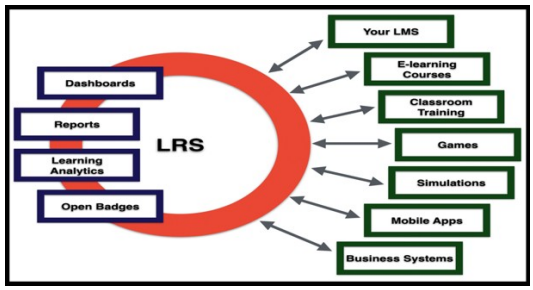


Figure 3: Learning Locker xAPI-Enabled Communications Across Different Devices

Hardware

In order to further demonstrate that local GIFT servers can operate in conjunction with modern VR and IOT devices while performing adaptive training, a variety of hardware was chosen with cost and existing available equipment being the main driver in selection. Not included are the PCs and peripherals necessary to run GIFT locally, and powerful enough to run VR-capable applications.

- Vive VR System: https://www.vive.com/us/
- 3-Watt, 8-Ohm Single Cavity Mini Speakers
- ¼-Watt, 470-Ohm Resistors
- Jumper Wires for Arduino

- Arduino Uno Microcontroller
- Arduino Network Shield
- 8,000 RPM Micro DC Motors
- 40V, 600mA, 300MHz, 625mW Transistors
- Red and Green 6-13V LED Diodes
- 10uF, 50V, 105c Capacitors
- Solderless PCB Breadboard
- Pulse Sensors
- ¼-Watt, 1k-Ohm Resistors
- 1000uF, 25V Capacitors

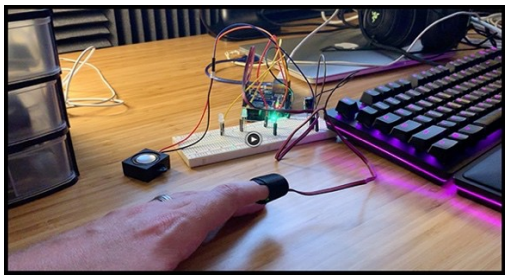


Figure 4: Arduino Hardware Components

The hardware listed above resulted in a combined system that was capable of monitoring a trainee's pulse rate tracking "health readiness," sending Go/No-Go visual and audio cues, sending indicator vibrations at different strengths through haptic motors attached to a trainee, all while performing a scenario in a desktop or VR Unity simulation tracking masteries referenced in CASS being monitored and managed by GIFT.

Methodologies

This section further explains the configurations, interfaces between disparate systems, and the experimental framework that was created as a result of this paper's effort.

GIFT Installations, Configurations, and Modifications

It is assumed that the reader has a working knowledge of the GIFT software suite in order to fully understand this section. Special attention is given to the new areas of study that this paper explores, but only brief mention is given to basic GIFT topics. For more information, the reader may refer to GIFT documentation at www.gifttutoring.org, or How-To YouTube videos at https://www.youtube.com/channel/UCWtI_V8f2mN5XD6h2lCjsAA.

Prerequisites to this section include having downloaded GIFT 2019-1, fully having configured GIFT server, and having configured network communications and hosting to the point of understanding the reader's system being localhost vs. hosted online at a specific IP address or DNS web address. The authors wished to fully demonstrate GIFT's interoperability with the distributed online world, and thus used fully-hosted online servers for all following development described in this paper.

A key concept to understand about configuring GIFT to be part of a distributed system is the nature of ActiveMQ / MQTT network communication. GIFT maintains its own ActiveMQ server on the machine it is running on, as

well as being able to communicate with other non- centralized network servers through many different methods such as RESTful web service calls or custom ActiveMQ/JSON messages. Fully explaining security, safety critical messaging, and ActiveMQ server configurations is beyond the scope of this paper, but the reader is encouraged to visit http://activemq.apache.org for further information. Of important note for this paper topic is the knowledge that the authors setup and configured another Mosquitto (MQTT) server to enable distributed communications.

CASS Server Installations and Database Entries

At the time of this writing, the authors are in the process of installing a CASS Server, the code of which is accessible by following links in the CASS Developer Guide here: http://devs.cassproject.org/index.html, on an Amazon Web Service (AWS) Elastic Compute Cloud (EC2) instance. While the instance hosting this new CASS server was not completed at the time of this writing, the authors are in the process of finalizing the configurations and will be offering connection information to the community as an additional test CASS server based on the most-current GitHub code base.

During many programs, such as ADL's Total Learning Architecture (TLA) initiative, various entities have interfaced with a CASS server and entered in various maturities of frameworks. One such entry, for instance, resulted in Army SMEs creating a custom framework for use in a Fort Benning experiment in late 2018. By enlisting subject matter experts to populate a CASS database with relevant competency relationships, an ITS such as GIFT becomes enabled to read and analyze these relationships to better-adapt training content tagged with similar metadata. The authors created a faux-framework for use during this paper's efforts at https://dev.cassproject.org/api/data/schema.cassproject.org.0.3.Framework/ba049c98-0d69-4fc3-96e1-931b90035fe3 which can be accessed in a GIFT course through a RESTful API call. CASS servers return data about frameworks and competencies in JSON formats.

By linking a GIFT Course Property menu item to the CASS database information, it became dynamically possible to populate a GIFT course that read in competencies that the course could then be linked to. For instance, a competency of "Attach Biometric Smart Clothing" <Broadens> a competency of "Sync Biometric Smart Clothing to Network." In reverse, the 2nd competency in the previous sentence <Narrows> the 1st competency according to CASS vocabulary. Other examples of CASS relationships include <Equivalent to>, <Requires>, <Is Enabled By>, <Is Related To>, and <Desires>. Every competency can be related to every other competency within a framework, and even to other framework's competencies as well.

The GIFT code enabling these CASS properties to be linked to a GIFT course will be made available to the public community in a following release of GIFT, and also made live in CLOUD GIFT at https://cloud.gifttutoring.org when ready for release by the development team.

This paper's effort concluded with a GIFT course reading information in from a CASS database and displaying it in a read-only fashion, but not storing any information permanently as exact specifications on how to integrate GIFT and CASS are still being discussed by the GIFT community.

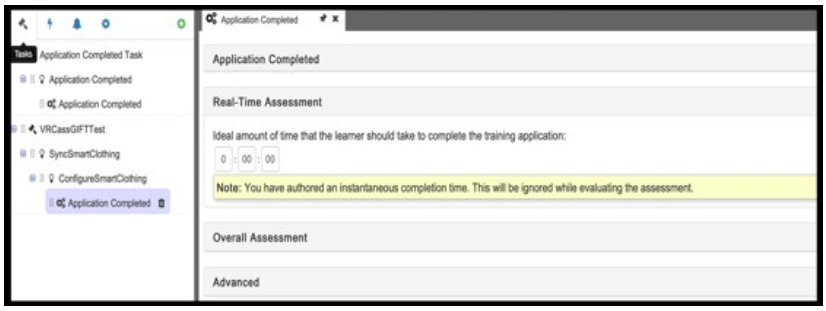


Figure 5: GIFT Unity Course Object Assessment

The live Unity simulation was also coded to read in CASS database information dynamically over the internet, and during the scenario was able to temporarily store "competency progression" as goals in the virtual world were completed, completing the GIFT-CASS-VR distributed system proof as all communication was handled through

an intermediary and agnostic Mosquitto server online at https://test.mosquitto.org/ (no 2 systems in this paper's effort were localhost, viewing all IOT devices together as "one system").

Virtual World Scenario in Unity

The Unity level was built primarily using existing art assets or low-cost acquisitions from the Unity Asset Store. It is recommended that if the reader wishes to integrate a Unity application with GIFT for the first time that the WebGL version of a Unity build be used according to existing integration instructions in the GIFT documentation and Course Object meant for such interactions.

The authors treated this paper effort's scenario as an external application, however, as demonstrating GIFT in a distributed environment was a desired outcome. The virtual world scenario can operate on any PC capable of running a basic Unity application, and communicated to the distributed Mosquitto server and the online CASS test server through the UnityWebRequest Unity library with which REST calls were made. For further information on enabling stand-alone Unity scenarios to communicate with non-local systems, the reader is encouraged to reference https://docs.unity3d.com/ScriptReference/Networking. UnityWebRequest.html.

The Unity scenario can be run in Desktop or VR modes that Unity allows for with first class citizen libraries. The scenario was also coded to display identical CASS database competency information in the Heads-Up Display (HUD). Also included in the HUD were representational self-health and squad-health biometric symbols, each of which represented the trainee in the lesson and simulated squad members, respectively. By allowing the trainee to see both the virtual world complete with buildings, non-player characters, equipment, furnishings, weather effects, while simultaneously displaying HUD information with "extrasensory" information about the environment, the authors wished to demonstrate not only virtual reality but what an augmented reality system could, in the future, begin to take the form of when integrated with GIFT.

The GIFT course operates in parallel with, and preferably launches, a virtual world scenario similar to how interfaces between GIFT and VBS are utilized. Any course training enabled through (and constrained by) virtual reality and/or any game engine can be constructed through using GIFT Course Objects and adaptive training measures. Examples include Real Time Assessments, Tasks, and Conditions that can be authored using the GIFT Authoring Tools (GAT) that respond to and inject events into the virtual environment scenario. Thanks to the open API design of both systems, training within a course topic and mix of competencies read in from CASS can be communicated from and to each of the discussed systems to great effect.

Shown below in Figure 6 is an example of how independent systems can communicate with a virtual world scenario. By listening to messages from the Mosquitto server, a simple HUD can listen for changes in self or squad health, offer visual indicators of masteries waiting to be acquired or already achieved, and offer advice to the trainee at a level the instructor or a GIFT course deems to be most-valuable to the training experience.

Figure 6: Unity HUD with CASS and Biometrics

IOT Devices and Communications

The complete setup instructions to configure an Arduino microcontroller with breadboards, custom voltages, wiring, and IOT devices is beyond the scope of this paper, but if interested, the reader is encouraged to reference the following online documents for instructions:

- Speakers:https://www.deviceplus.com/how- tos/arduino-guide/entry019/
- LEDs:https://www.deviceplus.com/how- tos/arduino-guide/entry_002/
- Long_Range_Communication:https://www.de viceplus.com/how-tos/arduino-guide/arduino- long-range-communication-tutorial-lorenz- shield/
- Pulse_Sensor:https://pulsesensor.com/pages/in stalling-our-playground-for-pulsesensor- arduino

With the prototype system libraries, hardware listed in the above 2.6 Hardware section, and GIFT-CASS-VR system operational, the same network paradigm of minimal Mosquitto messaging was used to link in IOT devices to the software suite. The authors began by connecting IOT devices to each other and a microcontroller, all of which are traditionally very limited in all of power, range, "disk space," APIs, and capabilities, and then connected these devices to network/long range communication capabilities. After integration efforts inspired largely by open source communities, the authors were able to have all IOT devices managed through a single microcontroller, report their sensor data, and respond to system messages passed-to and received-from the Mosquitto server.

While for this paper's effort the authors only managed the simplest of message passing between the IOT devices and the rest of the system, the primary goal of proving interconnectivity of a system of systems with GIFT acting as the ITS was satisfied. In addition, communication with parallel virtual environments (eventually formally multiplayer instead of simulated) and the distributed nature of the system as a whole was satisfied as a secondary goal. The authors look forward to future discussions with the community in these areas of interest.

Team Training Perspectives

The framework described thus-far resulted in a prototype system built with multiplayer and squad perspectives in mind. By creating a system in a distributed network with no (or at the very least, reconfigurable) single point of communication failure, a system in which nodes can be stood up-or-down on demand has had its foundation formed. Using principles and paradigms of load balancing made commercially-ready from companies such as AWS, nodes such as GIFT can better recover from a downtime perspective upon degraded performance being detected (a sensor slipping off, network communications being jammed, etc.). This also means a plug-and-play smart IOT uniforms could be developed and switched out with minimal interruption, or virtual scenarios dynamically configuring themselves to detect squad data for each team member and simulating or removing entities based on scenario configuration settings. With these distributed system paradigms in mind, frameworks such as the prototype presented in this paper form the basis for managing squad training in scenarios with ITSs (GIFT) and competency systems (CASS) in an active role, with completely variable sensor data and number of active players being switched out as training needs dictate.

Enabled Experiment Frameworks

The authors wish to provide the results of this effort back to the community as a nonprofit effort and will work with the GIFT team and community to determine the best path forward in this regard.

Some options include formal delivery of GIFT code updates to builds pushed out to CLOUD GIFT, specifically those updates relating to CASS database interactions. Other options include experimental branches that the community can request specific access to, or downloadable builds and configurations that enable capabilities based on individual or organizational needs.

As a prototype volunteer effort, the authors will work to the best of their ability to discuss any GIFT team and community interest to provide access to the experimental framework that can enable future research in the areas of:

- GIFT communication with LMS, LRS, or Competency (CASS) systems
- Competency metadata incorporation into GIFT courses
- Biometric and IOT device status monitoring in a GIFT course
- Virtual reality scenario creation and integration into a GIFT course
- Mixing 1-to-1 GIFT course and virtual world scenarios, 1 GIFT-course and single-player- team-simulation courses, and n-GIFT courses to n-virtual world scenario experiments in parallel

Conclusions and Future Research

The authors set out to perform a software engineering feasibility study that GIFT would be able to act as a centralized ITS in a decentralized system of systems. Through combining GIFT, CASS, Unity VR scenarios, and IOT devices, the authors have shown how GIFT's open API communication protocols allow for efficient integration into larger-scale training systems. Of specific interest to the authors is future involvement with systems such as CASS that enable, among other capabilities, a system of standards with which to track relevant learner skills and metadata to allow for improved adaptive training. Combined with an LRS and/or LMS, the authors hope to continue these paper's efforts to improve adaptive training by incorporating modern technologies and extending GIFT course functionality to include even more capability to handle simultaneous team member squad training.

References

All references for this paper are listed in-line within the content as all references were online resources.

Authors Biography

Zach Heylmun *graduated from the University of Florida with a degree in Digital Arts and Science engineering and supports SSI on the GIFT program as an as-needed consultant for the past two years. After graduation, he worked for Lockheed Martin on low-level, high performance graphics as well as virtual reality rendering for flight simulation and training. Since starting his own company, Voidstar Solutions, as well as helping to form Synaptic Sparks, a 501c3 charity dedicated to STEM education, he has worked with a wider variety of technologies. Through a combination of efforts, both for- and non-profit, he has worked on web technologies, mobile applications, and server infrastructures.*

Mike Kalaf *has over 30 years of Modeling, Simulation and Training leading large scale efforts leveraging cutting edge technology. Mike has worked in the commercial and military aviation, training and simulation business. In his most recent efforts, he has been leading new opportunities applying front end modeling, simulation and analysis. Mike has led several programs integrating "state of the art" technology and delivering highly successful technology and business innovation. Mike has been collaborating with educational organizations and exploring conceptual frameworks, platforms and business models to transform our current system and elevate the performance and quality. He is involved with the University of Central Florida's College of Education on a unique system of teacher training via classroom simulators. These projects fit well to advance science, technology, engineering and mathematics learning to lay the groundwork for a new generation of engineers and scientists. Mike volunteers his time to numerous education organizations including serving as a board member for the Central Florida STEM council and the Seminole County Public Schools Foundation. Mike's formal education includes an earned Mechanical Engineering degree from Rochester Institute of Technology, RIT.*

Christopher Meyer *brings a breadth of leadership experience and technical knowledge to the team. And, most recently, Christopher has supported the GIFT program for two years under the most current contract. He received his Bachelor and Master of Science degrees in Computer Science from Kansas State University, also receiving minors in Economics and Modern Languages, and studied abroad for a year during a tour in Japan at Chukyo University dedicated to the specialized study of Artificial Intelligence. After completing traditional education phases, Chris was employed at Lockheed Martin for 10*

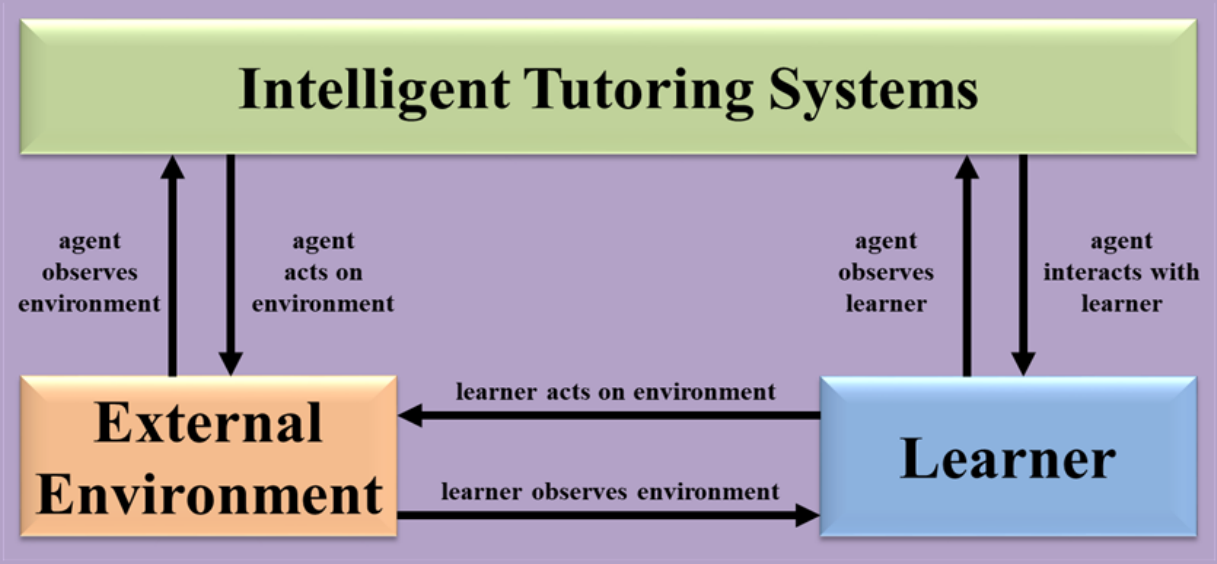
years working hand-in-hand with representatives from the Departments of Defense, Health and Human Services, Energy, and Education to assist in the creation of solutions to solve challenges at a national level. Having now co-created his own business segment, Chris enjoys utilizing entrepreneurship, international experience, leadership knowledge, and his own engineering skills alongside his peers to advance world technology, health, and opportunity efficiently and responsibly.

***Christofer Padilla** graduated from the University of Central Florida with a Bachelor of Science in Computer Science where he researched creating a usable concurrency library for submission to the Boost C++ libraries. He is currently employed by Dignitas Technologies, after completing a two-year internship as a developer on the GIFT program through two major releases. His contributions include supporting the development effort on the Real-Time Assessment Editor and the Course Authoring Tool. Christofer's early passion for software engineering has garnered him experience across many domains including virtual reality systems, computer graphics, web frameworks and intelligent tutoring systems.*

***Lucy Woodman** has recently graduated from Seminole State College with a Bachelor of Science in Information Technology. Lucy has supported Synaptic Sparks for one year during a successful internship and transitioned to be a major supporter of big data services within SSI in December of 2018. Lucy is a certified Amazon Web Service specialist and is currently studying to obtain further AWS certifications in System Architecting. Lucy also supports the team by providing research and development support in new fields of network and social technology.*

Proceedings of the Seventh Annual GIFT Users Symposium

GIFT, the Generalized Intelligent Framework for Tutoring, is a modular, service-oriented architecture developed to lower the skills and time needed to author effective adaptive instruction. Design goals for GIFT also include capturing best instructional practices, promoting standardization and reuse for adaptive instructional content and methods, and technologies for evaluating the effectiveness of tutoring applications. Truly adaptive systems make intelligent (optimal) decisions about tailoring instruction in real-time and make these decisions based on information about the learner and conditions in the instructional environment.



The GIFT Users Symposia began in 2013 to capture successful implementations of GIFT from the user community and to share recommendations leading to more useful capabilities for GIFT authors, researchers, and learners.

About the Editor:

- ***Dr. Benjamin S. Goldberg*** *leads adaptive training research at the U.S. Army Combat Capability Development Command – Solider Center and is a co-creator of the Generalized Intelligent Framework for Tutoring (GIFT).*

Part of the Adaptive Tutoring Series

9 780997 725766