

# Proceedings of the Twelfth Annual GIFT Users Symposium

May 2024  
Orlando, Florida  
and Hybrid



**GIFT**

*Edited by:*  
**Anne M. Sinatra**

**Part of the Adaptive Tutoring Series**

**Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)**

**Proceedings of the 12<sup>th</sup> Annual  
Generalized Intelligent Framework  
for Tutoring (GIFT)  
Users Symposium  
(GIFTSym12)**

*Edited by:  
Anne M. Sinatra*

APPROVED FOR PUBLIC RELEASE

**Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)**

Copyright © 2024 by the U.S. Army Combat Capabilities Development Command – Soldier Center.

**Copyright not claimed on material written by an employee of the U.S. Government.  
All rights reserved.**

**No part of this book may be reproduced in any manner, print or electronic, without written  
permission of the copyright holder.**

*The views expressed herein are those of the authors and do not necessarily reflect the views of the U.S. Army  
Combat Capabilities Development Command – Soldier Center*

Use of trade names or names of commercial sources is for information only and does not imply endorsement  
by the U.S. Army Combat Capabilities Development Command – Soldier Center.

This publication is intended to provide accurate information regarding the subject matter addressed herein. The  
information in this publication is subject to change at any time without notice. The U.S. Army Combat  
Capabilities Development Command – Soldier Center, nor the authors of the publication, makes any  
guarantees or warranties concerning the information contained herein.

Printed in the United States of America  
First Printing, August 2024

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

International Standard Book Number: 978-0-9977258-6-5

*We wish to acknowledge Alexandra Lutz for her effort in collecting the book chapters, and putting the book in  
format.*

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

# CONTENTS

<b>From the Editor</b> .....	<b>v</b>
<b>Theme I: New GIFT Features and Applications</b> .....	<b>9</b>
The GIFT Architectural and Features Update: 2024 Edition .....	11
<i>Nicholas Roberts, and Benjamin Goldberg</i>	
Integrating Mixed Reality, Physical Simulators, and Adaptive Learning: A Use Case for Implementation.....	27
<i>Michael Cambata, Thomas Lenz, and Randall Spain</i>	
<b>Theme II: Competency Frameworks</b> .....	<b>43</b>
Developing a Squad Competency Framework in STEEL-R.....	45
<i>Grace Teo, Michael King, Jennifer Solberg, Benjamin Goldberg, Gregory Goodwin, Meghan O'Donovan, and Clifford Hancock</i>	
Multimodal Measures for the Integration of Metacognitive Teamwork Processes During Simulation-Based Training.....	55
<i>Megan Wiedbusch, Ryan P. McMahan, Anne M. Sinatra, Benjamin Goldberg, Lisa N. Townsend, Joseph J. LaViola Jr., and Roger Azevedo</i>	
<b>Theme III: Artificial Intelligence Applications</b> .....	<b>67</b>
Leveraging TCAT for Advanced Team Communication Analysis and Performance Assessment in GIFT .....	69
<i>Randall Spain, Wookhee Min, Nicholas Roberts, Vikram Kumaran, Jay Pande, and James Lester</i>	
Large Language Models and Their Implications for Conversational Tutors and GIFT.....	77
<i>Vasile Rus</i>	
Integrating Machine Learning Models and GIFT.....	87
<i>Andy Smith, Randall Spain, Nicholas Roberts, Jonathan Rowe, Bradford Mott, and James Lester</i>	
<b>Theme IV: Applications of GIFT</b> .....	<b>95</b>

A Synthetic Training Environment for Assessing Changes in Team Dynamics with the Generalized Intelligent Framework for Tutoring .....97  
*Scotty D. Craig, Kevin Gary, Jamie C. Gorman, Vipin Verma, and Robert LiKamWa*

Enhancing Data Science Courses Pedagogy through GIFT-Enabled Adaptive Learning Pathways.....105  
*Fadjimata I. Anaroua, Qing Li, and Hong Liu*

**Theme V: Exercises and Experiential Learning.....117**

STEEL-R in Multinational Joint Training Exercises (STEEL-Rx).....119  
*Aaron Presnall, Biljana Presnall, and Benjamin Goldberg*

Automated Scenario Generation to Support Competency-Based Experiential Learning in GIFT .....127  
*Andy Smith, Randall Spain, Wookhee Min, Anne M. Sinatra, Jonathan Rowe, Bradford Mott, and James Lester*

Implementing a Longitudinal Performance Comparison Interface for Improved After Action Review in Experiential Learning.....137  
*Caleb Vatal, Naveeduddin Mohammed, Nicholas Roberts, Benjamin Goldberg, and Gautam Biswas*

# FROM THE EDITOR

**Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)**



## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

Welcome to the Proceedings of the 12<sup>th</sup> Annual GIFT User Symposium! This year we are celebrating 12 years of GIFT Symposiums and have accepted 12 papers for publication. All of the presentations that occurred at GIFTSym12, and the papers in this volume show the versatility of the Generalized Intelligent Framework for Tutoring (GIFT), and the work that is being done with GIFT.

GIFT is an open-source intelligent tutoring system (ITS) architecture that is freely available online at [GIFTtutoring.org](http://GIFTtutoring.org). There are both Cloud and Downloadable version of GIFT. GIFT has been developed with multiple goals in mind including supporting ITS research, and simplified creation of ITSs and Adaptive Instructional Systems (AISs).

Our fantastic team, and our program committee did a great job supporting the development of GIFTSym12, reviewing papers, and assisting with the facilitation of the event this year. We want to recognize them for their efforts:

- **Benjamin Goldberg**
- **Gregory Goodwin**
- **Tamara Griffith**
- **Michele Myers**
- **Alexandra Lutz**
- **Randall Spain**
- **Lisa N. Townsend**

The themes for this year's GIFTSym include:

- New GIFT Features and Applications
- Competency Frameworks
- Artificial Intelligence Applications
- Applications of GIFT
- Experiential Learning and Exercises

The editor and program committee would like to thank all of the contributions and authors on the papers in this proceedings. We also would like to thank all those who contributed to GIFTSym in previous years. The feedback, lessons learned, suggestions and research that have been provided from GIFTSym through the years have been important in the development of GIFT.

We would also like to encourage readers to visit the documents tab on [www.GIFTtutoring.org](http://www.GIFTtutoring.org). Proceedings from each year of GIFTSym as well as the Design Recommendations for Intelligent Tutoring Systems book series are available for free on the documents tab. We also encourage users to sign up for a free GIFTtutoring.org account so that they can receive GIFT news, access to the user forum, and access the GIFT software.

Thank you for 12 great years of GIFTSym!

Anne M. Sinatra, Ph.D.  
GIFTSym12 Chair and Proceedings Editor

**Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)**



# **THEME I: NEW GIFT FEATURES AND APPLICATIONS**



# The GIFT Architectural and Features Update: 2024 Edition

Nicholas Roberts<sup>1</sup>, and Benjamin Goldberg<sup>2</sup>

Dignitas Technologies<sup>1</sup>, U.S. Army Combat Capability Development Command (DEVCOM) – Soldier Center<sup>2</sup>

## 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 regarding the architectural “behind the scenes” work and how the GIFT project is addressing user requirements suggested in the previous GIFTSym11 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 previous GIFT Symposium architecture updates (Brawner & Ososky, 2015; Ososky & Brawner, 2016; Brawner et al., 2017; Brawner & Hoffman, 2018; Brawner et al., 2019; Goldberg et al., 2020; Hoffman et al., 2021; Goldberg et al., 2022) this version highlights new tools and feature requests accomplished over the latest development cycle. The feature requests and architectural improvements are derived from two primary sources: (1) recommendations from symposium papers and other sources collected across the GIFT user base, and (2) stakeholder interactions linked to capability and project needs. The features are organized into logical sections within this update and cover modifications across all core modules operating within GIFT.

## 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 (Sottolare et al., 2012). First, we would encourage you to simply get started, as the tools and example courses have been designed to assist users in exploring GIFT’s tools and methods for the purpose of creating Adaptive Instructional Systems.

If you struggle with any individual aspect of the system, the team has produced short “how to” videos to help around the sticking points. There are now many 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 have remained unchanged.

Outside of the introductory materials and tutorials available in GIFT, there is also developer support through detailed documentation and active help forums. The GIFT user community is 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 community. The forums, at the time of this writing, have over 1700 postings and responses.

Documentation has been made freely available online at <https://gifttutoring.org/projects/gift/wiki/Documentation>, with interface control documentation available at

[https://gifttutoring.org/projects/gift/wiki/Interface\\_Control\\_Document\\_2023-1](https://gifttutoring.org/projects/gift/wiki/Interface_Control_Document_2023-1), and a developer guide available at [https://gifttutoring.org/projects/gift/wiki/Developer\\_Guide\\_2023-1](https://gifttutoring.org/projects/gift/wiki/Developer_Guide_2023-1). These documents are updated for each software release. In this paper, we would also like to highlight the available instructions for hosting your own Amazon Web Services (AWS) instance ([https://gifttutoring.org/projects/gift/wiki/Amazon\\_Web\\_Service\\_Install\\_Instructions](https://gifttutoring.org/projects/gift/wiki/Amazon_Web_Service_Install_Instructions)).

## GIFT Development and Release Strategy

There are two GIFT instances available to everyday users, GIFT Cloud and GIFT Desktop. GIFT Cloud follows an every-Friday system update schedule when relevant updates are ready from the engineering team. For the desktop version, we maintain a 12 month or less release cycle, with a recent regression-tested release in November 2023. To support experimentation, intermittent extensions of the core GIFT baseline are performed to facilitate data and interaction requirements based on specific research questions of interest. These are performed on an “as needed” basis, and often serve as the feature extensions included in the next public-release. Adjustments to the release strategy will be considered as more agile software development approaches are being applied at the organizational and enterprise level. As a member of the community, if you see a feature in the cloud release which you would like to use locally, simply ask.

## Cloud General Reporting

GIFT Cloud (see Figure 1) has been running continuously for the last eight years via Amazon Web Services (AWS). The cloud instance is kept online and updated in advance of the downloadable version, meaning that cloud 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 use of the Cloud version. It supports hundreds of simultaneous users for experiments. We are generally confident in the systems’ ability to stay up and cope with demand. The current limitations are that team training in a virtual environment and sensor-based interactions are not supported on the cloud instance, but that requirement will be addressed.

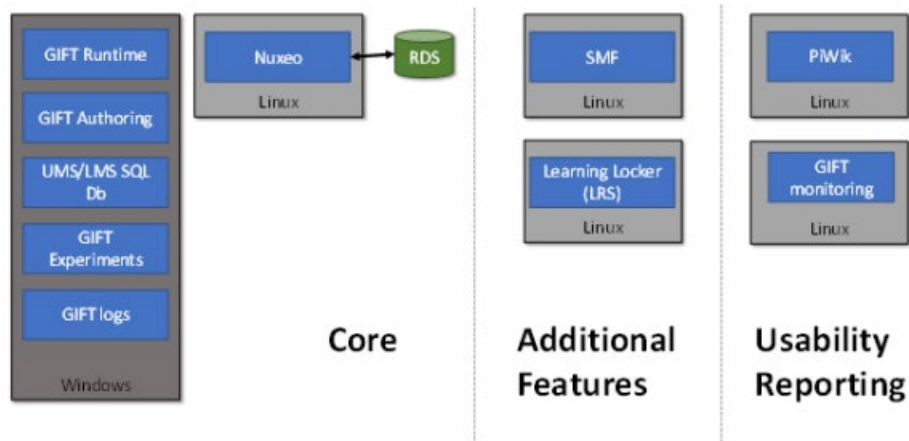


Figure 1. Simple Diagram Overview of GIFT Cloud Components

Behind the scenes, however, the re-tooling to move from a deployment version of development in a desktop instance to a cloud environment in production has been working well. For the remainder of the paper, we will cover the latest improvements added over the last development cycle.

## NEW GIFT FEATURES AND UPDATES

---

Since the last feature update from GIFTSym11 (Roberts et al., 2023), there have been multiple additions to the GIFT capability set. Each tool or method described in this section is now available in the latest public-facing open-source version of GIFT or on GIFT Cloud. Each new feature will be presented with information on the functions it supports and the system and data level dependencies to implement.

### Linux and Docker Support Updates

Support for running GIFT in Linux and Docker environments has been expanded since last year and is no longer considered an experimental feature as of the 2023-1 release. Scripts for running GIFT in Linux have been improved to reduce manual steps needed to perform the initial setup and now no longer require a separate patch from [gifttutoring.org](http://gifttutoring.org) to operate. The few dependencies that are still needed to run in Linux are downloadable from the [Download page on gifttutoring.org](http://gifttutoring.org) and can be placed in the GIFT/external folder to allow the installation script to perform the setup. In addition to these scripting changes, improvements have also been made to the codebase to better support a few common GIFT operations in Linux. If a course attempts to use a training application that does not support Linux, GIFT will now display an error message that clearly indicates which operating systems the given training application supports. GIFT's logic for playing back log files was also found to have issues with taking logs that were created on one operating system platform, such as Windows, and then running them on a different platform, such as Linux. This logic has been improved so that GIFT instances on Windows or Linux can play back logs from either platform.

Several improvements were also made specifically to support using Server mode within Linux environments, with the goal of allowing server instances of GIFT, like GIFT Cloud, to run on lightweight Linux virtual machines. For instance, a “stopGIFT” script has been added to allow all running GIFT processes to be quickly shut down, allowing for faster redeployments. The Java Management Extensions (JMX) settings in GIFT's launching scripts have also been improved to properly establish JMX endpoints for all of GIFT's processes for both Windows- and Linux-based servers. This allows Java management tools like VisualVM and JConsole to be used to debug and maintain servers. It also allows GIFT to produce usage metrics in Linux that can be monitored using management tools like Grafana, improving long-term maintenance and allowing server operators to more easily identify problems. Altogether, these changes allow Linux-based instances of GIFT to exercise greater control over GIFT's modules and applications, similar to what GIFT has supported in Windows since the initial introduction of Server mode. See Figure 2 for an example of GIFT being started using Ubuntu.

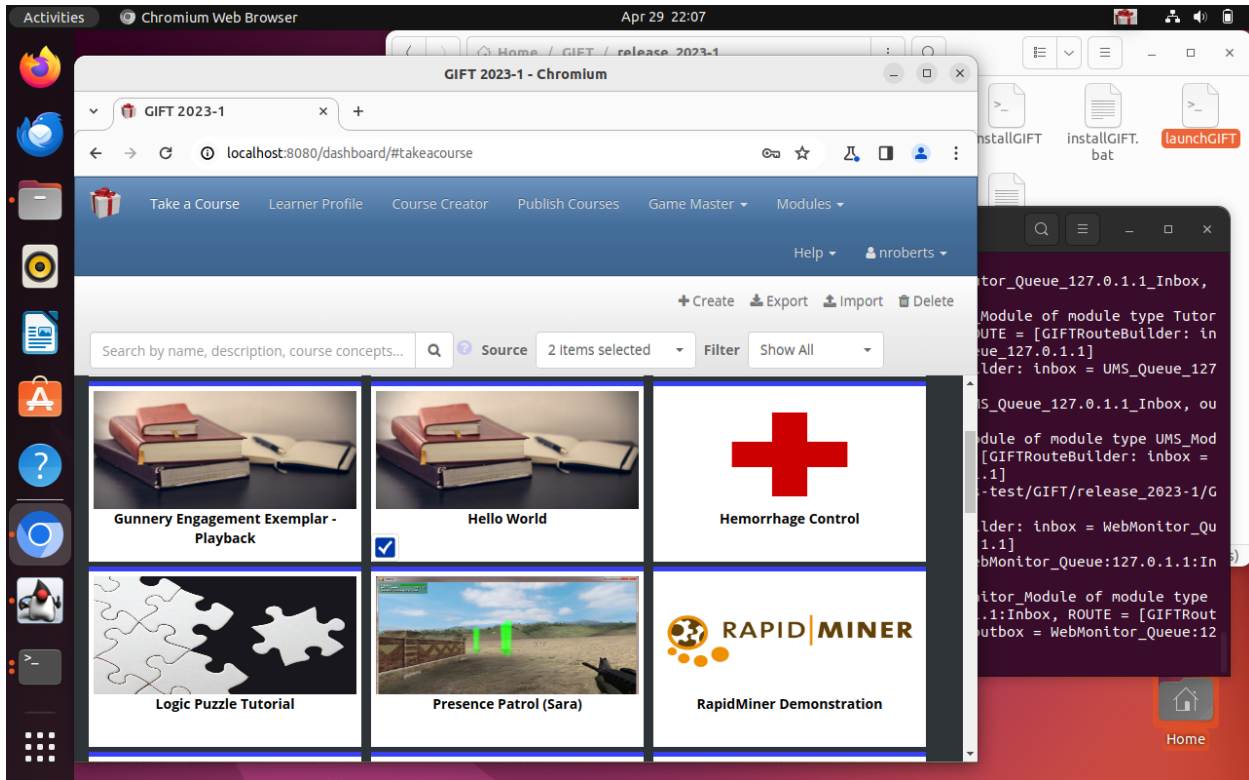


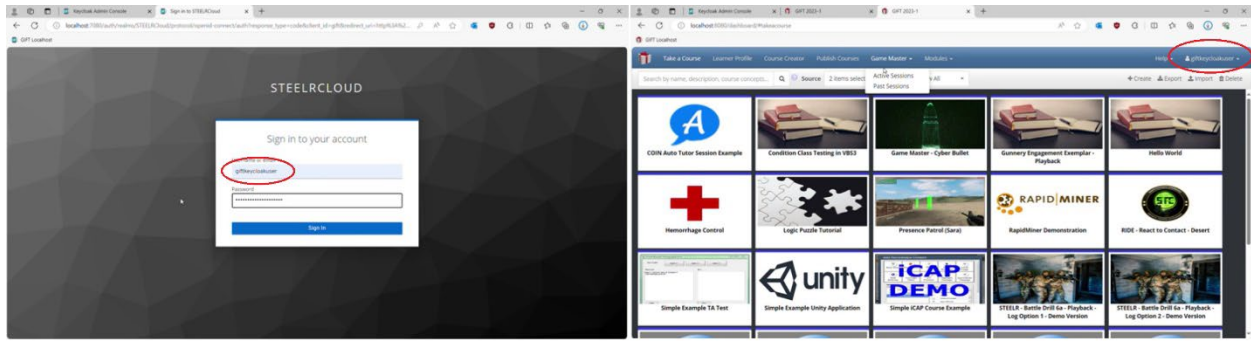
Figure 2. In the screenshot above, GIFT has been started in Ubuntu using the “launchGIFT” script, and its “Take a Course” webpage has been opened in Chromium to run a course.

## Keycloak and Single Sign-On Support

GIFT can now be configured to let users sign into its webpages using Single Sign-On (SSO) authentication credentials, allowing deployments to use existing authentication systems that are not directly managed by GIFT. GIFT has traditionally allowed users to sign into its webpages using their account credentials on [gifttutoring.org](https://gifttutoring.org), and while it could also be configured for users to sign in offline or using explicit predefined credentials, it previously did not allow for another authentication system to handle these sign-ins and bypass GIFT’s own login pages. With the latest release, however, the GIFT Admin Server (GAS) configuration files now provide settings that can be modified to allow GIFT to use other SSO-based authentication systems. Specifically, these configurations support SSO solutions that leverage OpenID Connect (OIDC), such as Keycloak. Keycloak, in particular, has had the most explicit support, since a special configuration file has been added to explicitly accept Keycloak OIDC installation settings. That said, several of these new configurations leverage generic security restraint rules from the Java Server API and thus could be used to support other authentication systems. If GIFT has been configured to use a SSO provider, it will display the login page from that SSO provider instead of the default login page. GIFT will also track a user’s username within the SSO provider, display it within GIFT’s webpages, and save it to appropriate database calls, ensuring that users from the SSO are still given separate workspaces when running in Server mode. Finally, if the SSO provider that GIFT is using also allows the same user to visit other websites, then signing into one of those websites and then visiting GIFT’s webpages will bypass any login screen that GIFT would normally present, allowing users to seamlessly navigate to and from GIFT if they have signed into the SSO at least once within the current browser session. An example of this can be seen in Figure 3.



## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)



**Figure 3.** In the screenshot to the left, GIFT has presented the user a login page from Keycloak where they have entered their SSO credentials. After signing in, the picture on the right shows their username within the SSO system rather than their username on gifttutoring.org.

### Experience Training Support Package (XTSP) Import and Export

Adding onto behavior introduced in the previous GIFT releases, GIFT’s course creator now features improved logic for importing Experience Training Support Package (XTSP) files and allows users to export certain modifications back to XTSP files. For some background, XTSP files are JavaScript Object Notation (JSON) files that define a structure of tasks, standards, and conditions that can be imported by GIFT’s Course Creator to create a Domain Knowledge File (DKF) for a training application course (Goldberg et al., 2023). This importing behavior has been improved to convert more XTSP elements into their corresponding counterparts in GIFT’s DKF schema. Notably, GIFT will now import triggers that are used to start or end XEvents within an XTSP file and convert them to start and end triggers for DKF tasks that are created from the same XEvents. It will also import any XTSP strategies that initiate “Actor intervention” or “Custom script” activities and convert them into their equivalent DKF strategy types. The course creator’s XTSP importing logic will also now display validation warnings if issues are found with converting XTSP elements into their DKF equivalents. Typically, these issues occur if the XTSP file does not match the version of the XTSP schema that GIFT understands. Previously, if GIFT encountered issues like these, it would make a best-effort attempt to convert the problematic XTSP elements, but it would not clearly tell the user why these elements were problematic. Now, if such issues are encountered, the course creator will display a window outlining the issues that were found, which can then be closed at the author’s discretion. This behavior does not interrupt the importing process or stop the best-effort conversion from happening, but it makes it more obvious when such issues occur in case the author has accidentally imported the wrong version of an XTSP file or in case there are validation problems with the XTSP that the author is not aware of.

In previous releases, a notable limitation of the import behavior was that changes that were made to a DKF imported from an XTSP file could not be written back to the same XTSP file later, but now this is no longer the case. When a DKF is created from an XTSP file, a copy of the original XTSP file is saved to the course folder. Changing certain objects within this DKF and then saving will update the XTSP file accordingly to “export” these changes back to the XTSP file, keeping the DKF and XTSP file roughly in sync. The newly imported start and end triggers are a good example of this; if task start and end triggers are modified in a DKF that was created from an XTSP file, then those modifications will be written back out to the copy of the XTSP file. The same is also true for places of interest, which means adding, modifying, or removing places of interest will affect the XTSP file’s overlays in a similar manner. One set of elements particularly of note regarding this exporting logic is a DKF’s condition inputs. The XTSP schema is intentionally open-ended in how it defines method inputs so that applications like GIFT can inject their own metadata, which GIFT uses to save condition inputs to an XTSP file. If a DKF created from an XTSP file has its condition inputs changed, then those changes are serialized to strings and written to the XTSP file. If GIFT later

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

imports an XTSP file modified in this manner, then it will import these condition inputs alongside the step measures, just like it imports other elements of the XTSP schema. It is worth noting that GIFT cannot generate an XTSP file from scratch using this exporting logic, and any elements that are not explicitly mentioned above will not be modified when exporting back to the XTSP. See Figure 4, below, for an example of data that was exported to an XTSP file using GIFT.

```
"teamSkillMeasures" : [ {
  "msrId" : 9,
  "msrUuid" : null,
  "msrTitle" : "Team Skill 1 Measure 1",
  "msrClass" : "individual",
  "msrType" : "teamwork-skill",
  "position" : 1,
  "weight" : null,
  "msrConditions" : [ "domain.knowledge.condition.AssignedSectorCondition" ],
  "evalMethod" : "Manual",
  "evalClass" : [ "Manual Observe" ],
  "dataSources" : null,
  "methodInputs" : [ "<?xml version='1.0' encoding='UTF-8' standalone='yes'?'>\n<input xmlns:ns2='http://GIFT.com/common' xmlns:ns3='http://GIFT.com/learnerActions'>
\n <AssignedSectorCondition>\n <teamMemberRefs>\n <teamMemberRef>BLUFOR Division Role</teamMemberRef>\n <teamMemberRef>BLUFOR Brigade
Role</teamMemberRef>\n <teamMemberRef>BLUFOR Battalion Role</teamMemberRef>\n <teamMemberRef>BLUFOR Company Role</teamMemberRef>\n
<teamMemberRef>BLUFOR Platoon Role</teamMemberRef>\n <teamMemberRef>BLUFOR Squad Role</teamMemberRef>\n <teamMemberRef>BLUFOR Fireteam
Role</teamMemberRef>\n <teamMemberRef>OPFOR 1A</teamMemberRef>\n <teamMemberRef>OPFOR 1B</teamMemberRef>\n <teamMemberRef>OPFOR
2A</teamMemberRef>\n <teamMemberRef>OPFOR 2B</teamMemberRef>\n <teamMemberRef>Admin Role</teamMemberRef>\n <teamMemberRef>OPFOR
Value='Activate Enemy Contact'</>\n <maxAngleFromCenter>45</maxAngleFromCenter>\n </AssignedSectorCondition>\n</input>\n" ],
  "defaultLevel" : null,
  "formativeCriteria" : [ ],
}
```

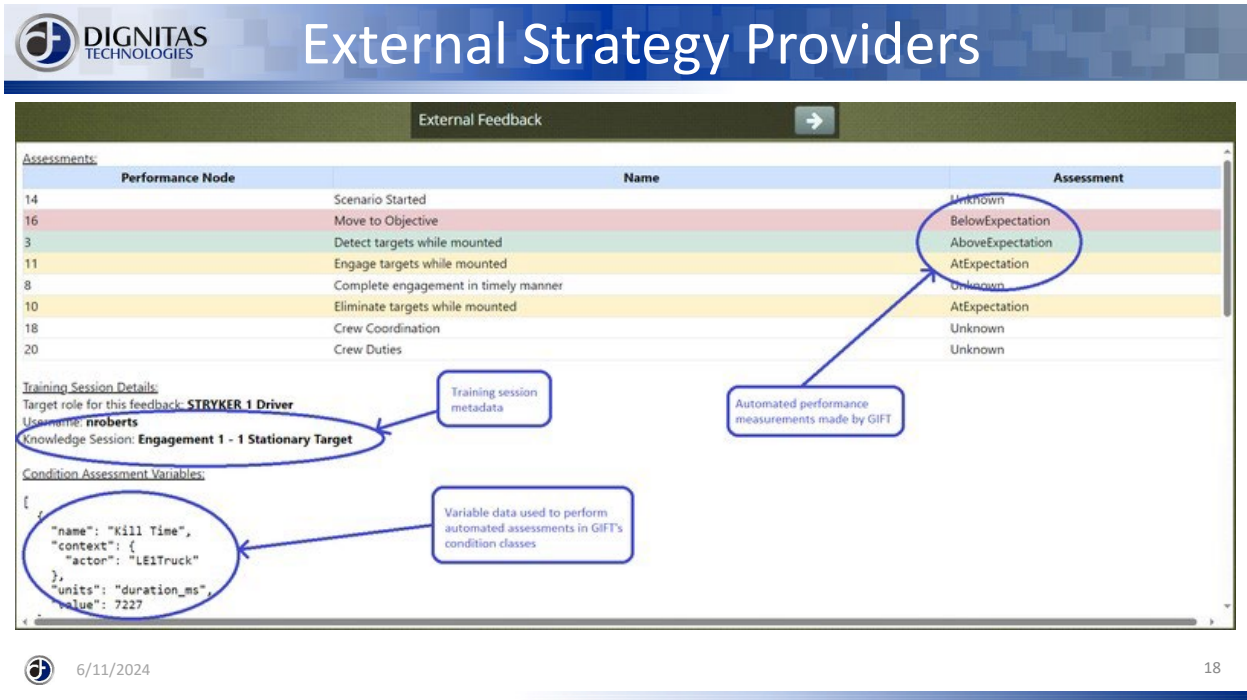
**Figure 4.** The snippet of an XTSP file shown above gives an example of data that was exported to an XTSP file using GIFT. In this case, an Assigned Sector condition was added to a concept in the DKF, which was then saved to the appropriate measure object within the XTSP file.

### External Strategy Providers

External strategy providers are applications that accept requests from GIFT to generate material to show during strategies within a DKF training session. GIFT can now be configured to register these external strategy providers, which lets course authors leverage them when creating instructional strategies. An external strategy provider is expected to provide a Representational State Transfer (REST) Application Programming Interface (API) endpoint that GIFT can use to send information about the running training scenario and then receive strategy material to display in response. This REST API endpoint can be registered to GIFT by updating the new “ExternalStrategyProviderURL” in GIFT’s common.properties configuration. The information that GIFT sends to the external strategy provider is a JSON message that includes which learners are participating in the training, which team roles from the DKF have been assigned, and how the learners are currently performing in all of the tasks that they need to complete. Additionally, a new capability has been added to allow GIFT’s condition classes to share their internal metrics and measurements along with this information. This allows external strategy providers to gather more detailed information about the context of a given assessment and how a condition arrived at the assessment that it did. As an example, using a combination of the Detect Objects condition, Engage Targets condition, and Eliminate Hostiles condition will produce several gunnery-focused metrics, such as the time that object detection began, when a target was first engaged, and how long it took to eliminate the target after it was identified. Ultimately, this gives external strategy providers the ability to generate custom content that fits the current state of the scenario, rather than using static content that is already defined in the DKF.

To demo this new feature, along with several of the remaining features covered by this paper, a new “Gunnery Engagement Exemplar - Playback” course has been included with the 2023-1 release as a showcase course. This course plays back a recording of a gunnery training scenario that was conducted using Virtual Battle Space (VBS). In this scenario, a gunner, commander, and driver were tasked with using a gun-mounted STRYKER vehicle to eliminate targets in 3 training exercises, each with their own DKF. This playback course makes use of an example external strategy provider that is packaged with the 2023-1 release. When each training exercise is completed during the playback, the course uses the example external strategy provider to display the current user’s name, the assessments that have been calculated for all of the DKF’s tasks and concepts, and a breakdown of the gunnery metrics, or “variables” that were calculated by

the DKF’s condition classes. An example of this can be seen below in Figure 5. Instructions for setting up and running this course can be found in [the 2023-1 release notes wiki page](#).



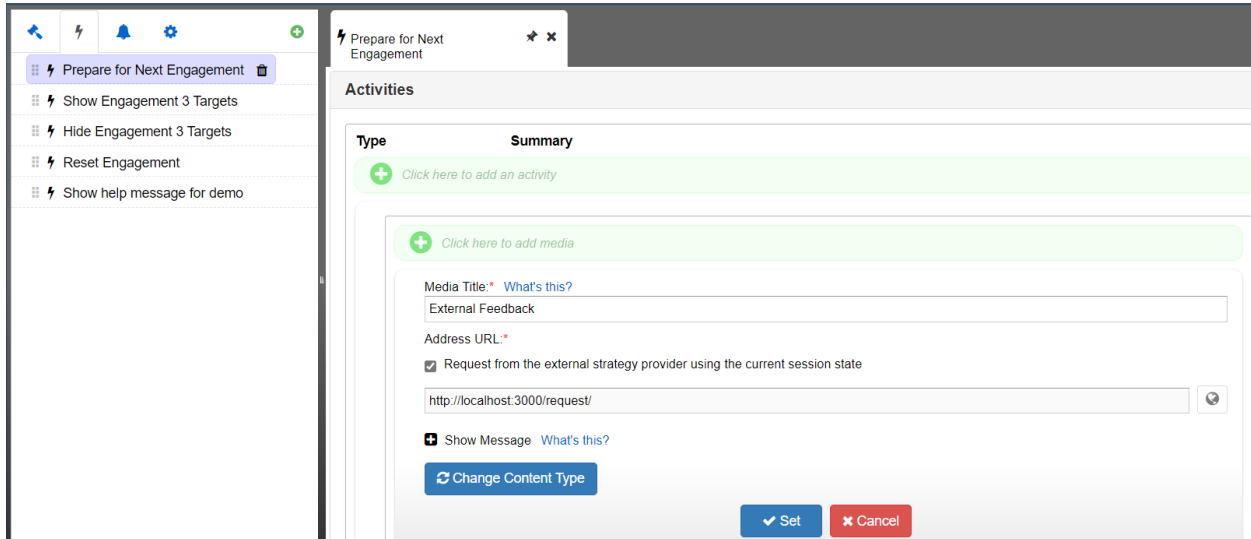
**Figure 5.** This is an example of a webpage that was generated from live training session data using the example eternal strategy provider that is packaged with the 2023-1 release. To the top right, assessment measurements are shown, while to the bottom left, the webpage displays metadata surrounding the training scenario as well as details surrounding condition metrics. In this case, the “Eliminate targets while mounted” concept used an Eliminate Hostiles condition, which produces the “Kill Time” metric that is displayed.

### Course Editor Updates

Coinciding with the new GIFT functionality that supports external strategy providers, GIFT’s course editor has also been modified to allow course authors to specify when exactly external strategy providers should be used within a DKF. If an external strategy provider has been registered to GIFT, then the course editor will display a new “Required from the external strategy provider using the current state” checkbox in a few specific areas to allow the user to choose when to request content from an external strategy provider. Specifically, this checkbox appears when authoring an instructional strategy using two activity types: “Present Media” and “Feedback”.

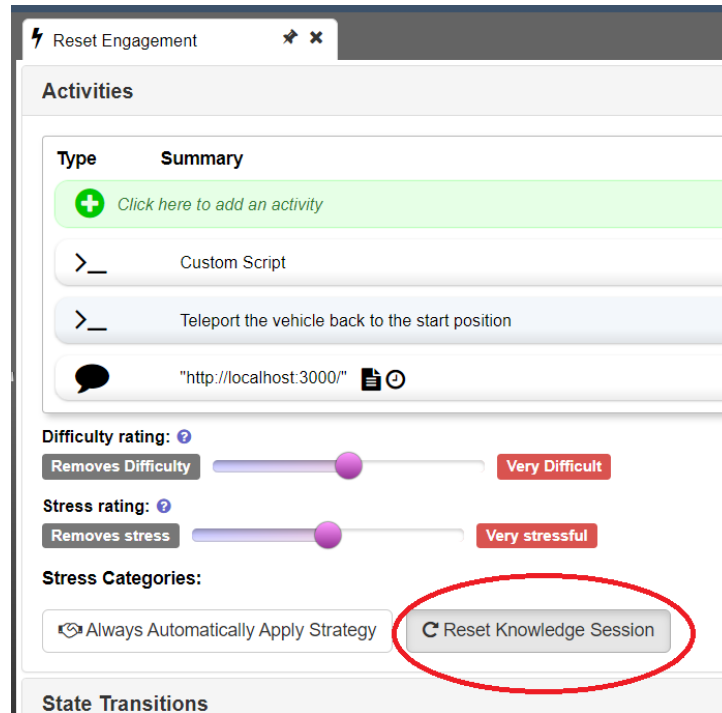
If the author chooses to use a “Present Media” activity and adds a “Web Address” item, this checkbox will appear, and if it is selected, then GIFT will send data to the external strategy provider when the activity is invoked and will expect the strategy provider to return the address of a webpage to display. Likewise, if the author chooses to use a “Feedback” activity and opts to “Present a Message” to the user, then the same checkbox will appear. If selected, this checkbox will again prompt GIFT to send data to the external strategy provider when the activity is invoked, though this time, it will expect a raw text message or HTML message to be returned by the REST call. In either case, the course will display either a webpage or a text message as soon as the external strategy provider has sent a response. The external strategy provider does not have to be running while these activities are being authored in the course creator, but if the author has selected the checkbox, then the external strategy provider is required to run when the course is executed by a learner.

If GIFT attempts to request strategy content from an external strategy provider and that provider is not running, then the course will terminate prematurely with an error. Figure 6, below, shows an example of the interface that has the option to request the webpage from an external strategy provider selected.



**Figure 6.** Here, the course author has created an instructional strategy to show a web page. The option to request the webpage from an external strategy provider is selected, so a URL pointing to the external strategy provider is specified for the author.

Similarly, the strategy editor in the course creator has also been given a new “Reset Knowledge Session” button. This allows course authors to specify that specific instructional strategies should reset a DKF training session back to its initial state. Unlike restarting the course entirely, this preserves learners within the training session and maintains their current team role assignments, skipping the team lobby screen. This is helpful for allowing course authors to restart a training exercise if the learners get into a state where they cannot proceed, and it can also be used to allow Observer Coach/Trainers (OC/Ts) to manually restart a training exercise from Game Master using buttons in the Scenario Injects panel. It should be noted that triggering a reset in this way only resets the state of the training session within GIFT itself and not within the training application, which means tasks and concepts within a DKF will reset to their initial states and assessments. As an example, if a learner moves to a new location within the training application and the session is reset, they will not be teleported back to their start location. Because of this, if the course author has clicked the “Reset Knowledge Session” button within a strategy, any activities that they add to this same strategy will be executed before the reset occurs and must complete first for the reset to occur. See Figure 7 for an example of the “Reset Knowledge Session” button. This gives the author the option to inject scripts or other environment adaptation activities to reset the scenario within the training application to its initial state before resetting the DKF training session. Additionally, any attempts at the DKF training session that end in a reset will still be saved and can be played back in Game Master, should an instructor or OC/T wish to review them later, and their performance will be recorded in xAPI statements if GIFT is connected to a Learning Record Store (LRS). These reset attempts will not impact the results shown in Structured Reviews or the summative scores that are shown in Game Master for the final attempt at the training session, so the learner’s overall experience over these attempts is tracked without affecting their final scores.

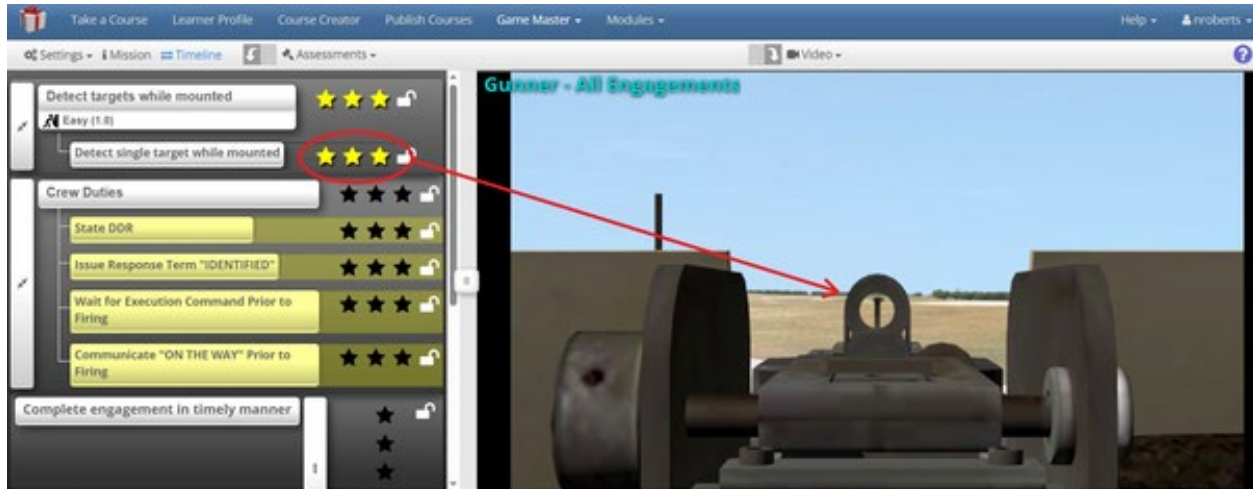


**Figure 7. The "Reset Knowledge Session" button has been selected with activities to teleport the vehicle containing the learners back to its start location. When the "Reset Engagement" strategy is invoked, the teleport will happen first before resetting the learners' tasks and assessments in the DKF.**

## Domain Module Updates

The logic that GIFT uses to manage DKF training sessions within the Domain module has been modified to better support courses that use multiple DKFs. Previously if a course contained multiple DKFs that each had more than one playable team role, then any learners would have to select their roles every time a new DKF training session began, even if the team roles were identical between all of the DKFs. Now, if a course uses multiple DKFs that share the same team roles, the roles that the learners selected for the first DKF training session are now remembered and used to automatically assign them to the same roles in any subsequent DKFs. This bypasses the team lobby screen entirely, allowing for learners to seamlessly move from one DKF to another during a team-based training exercise.

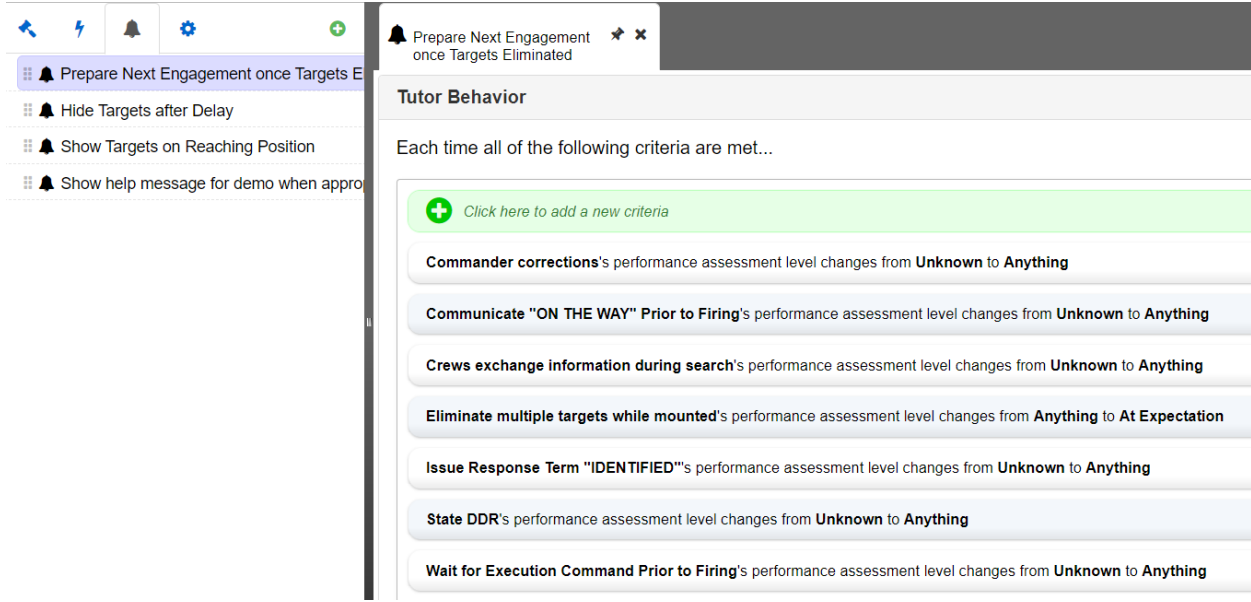
For vehicle-based training, the Detect Objects condition within the Domain module has also been modified so that a learner operating a gun mounted on a vehicle can properly trigger object detection events. During testing related to gunnery training in VBS, it was found that the Detect Objects condition did not produce assessments for any learners that were inside a vehicle, even though these same assessments would be produced normally if the learner was outside the vehicle. Further testing revealed that once the learner's player unit entered the vehicle, VBS would stop sending orientation data from the unit and, therefore, could not detect when the learner's orientation faced toward one of the objects that they were required to detect. To address this, the Detect Objects condition now first checks if the learner is present in a vehicle before performing its orientation calculations. If the learner is in a vehicle and that vehicle has a mounted gun, then the direction that the gun is facing is now considered the learner's orientation. If the vehicle does not have a mounted gun, then the facing direction for the front of the vehicle is considered the learner's orientation instead. With these changes in place, the Detect Objects condition will now properly produce an assessment if a learner aims at one of the target objects while operating a vehicle-mounted gun. See Figure 8 for an example of an AAR playback from Game Master that uses the Detect Objects condition.



**Figure 8.** This AAR playback from Game Master shows that a learner operating a mounted gun on a vehicle has triggered an Above Expectation assessment in the “Detect single target while mounted” concept, which uses the Detect Objects condition.

## Pedagogical Updates

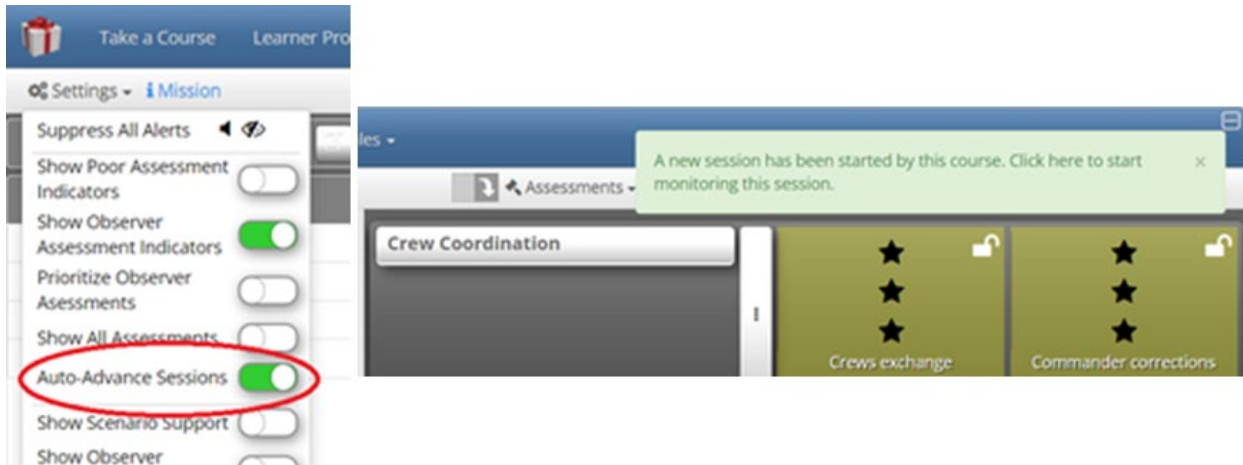
The pedagogical logic that GIFT uses to decide when to apply instructional strategies has also been improved to better address situations where an author has defined multiple criteria that must be met simultaneously to trigger a strategy. For context, a DKF determines when to invoke strategies by looking at the state transitions that are defined by the author, each of which defines one or more criteria that must be met in order to invoke one or more strategies. As part of the aforementioned gunnery training, one of the state transitions required an OC/T to provide manual assessments to several concepts within the DKF before moving on; however, testing revealed that this state transition would only be fired when the concepts were provided assessments in a particular order. This happened because of how the Pedagogical module was maintaining the “satisfied” state of criteria in between performance assessments. If the assessments were performed in an order that did not match the order in which the criteria were loaded, then any of the criteria that were out of sequence would have their “satisfied” states reset, preventing the transition from firing its strategies. This was fixed by improving the Pedagogical module logic to properly maintain the “satisfied” state in between performance assessments. Now, if a state transition has multiple criteria, it will fire its strategies once all of the criteria are concurrently satisfied. The key word here is “concurrently”. See Figure 9 for an example of criteria in the DKF. If one of the criteria was previously satisfied but then a new performance assessment is performed that un-satisfies it, then the state transition will not fire its strategies until that criterion is satisfied once again.



**Figure 9.** In this state transition, 7 criteria must be satisfied for the next engagement, i.e. the next DKF, to be started after eliminating all of the targets. With the new pedagogical behavior, these criteria can be satisfied in any order to start the next engagement, as long as all of them remain satisfied.

## Game Master Updates

Game Master has received several improvements mainly intended to allow OC/Ts to manage training sessions with multiple DKF exercises more easily and rapidly. The first of these improvements is that Game Master will now automatically move along to observe the DKF training session that is currently being executed if an active course contains multiple DKFs. Previously, if a DKF training session ended while Game Master was observing it, Game Master would remain on that session so that the observer could interact with it after the fact by switching to view it as a past session. This behavior worked well for courses with a single DKF, but it became problematic if multiple DKFs were present in a course, since moving onto the next one would require an OC/T to manually reload the “Active Sessions” list and select the next session. Having Game Master automatically switch to view the next DKF in the course avoids this problem and allows an OC/T to follow the current state of the course more easily. That said, in case an OC/T would prefer to control Game Master progression manually like with the old behavior, a new “Auto-Advance Sessions” setting has also been added to Game Master’s settings panel. This setting is turned on by default, and if it is turned off, then Game Master will remain on the most recently completed DKF until the OC/T chooses to move on, as it did previously. Rather than requiring the OC/T to reload the active session list, however, turning this setting off will also cause a notification to be shown to the OC/T if a new DKF training session is started while they are still looking at an old one. Clicking on this notification without dismissing it will allow the OC/T to immediately jump to the currently running DKF session, bypassing the need to return to the active sessions list while still giving the OC/T control over the behavior. Altogether, these changes make it easier for OC/Ts to work with courses containing multiple DKFs by making it faster and simpler to move through the sequence of DKFs. See Figure 10 for an example of the OC/T disabling the “Auto-Advance Sessions” setting.



**Figure 10.** If an OC/T disables the "Auto-Advance Sessions" setting shown on the left, then reaching the end of a DKF session while it is being viewed will display the message shown on the right if another DKF session is started. If this message is clicked on, Game Master will display the next DKF session, instead of remaining on the old session that just finished.

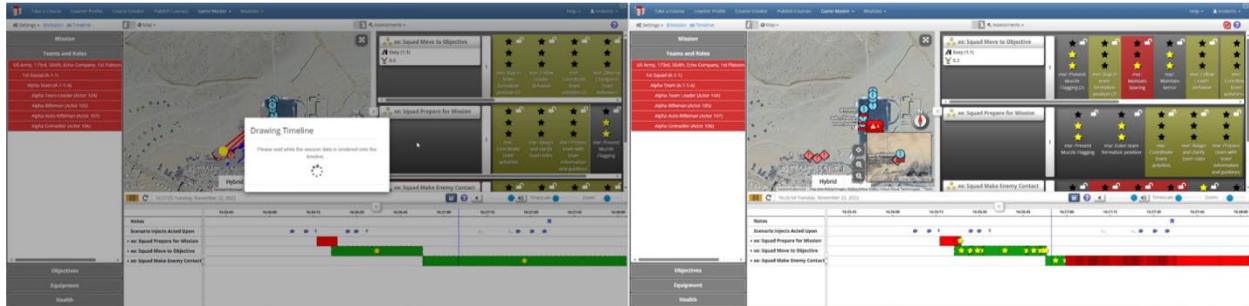
Coinciding with the new setting that was added to automatically advance through DKF sessions, Game Master will now save any modifications that are made to the settings in the settings panel. Previously, leaving Game Master at any point would reset any settings that had been adjusted, but now, any adjustments to the settings are recorded using browser cookies. If a user revisits Game Master after adjusting these settings, then they will be reloaded and put into effect immediately once the user returns to Game Master. This behavior has also been extended to the left and right panel selections in Game Master, so adjusting the panels that are shown will also be saved. This can be used to avoid seeing the map upon entering Game Master when it is not relevant, since selecting a different left panel will change the default panel that is shown upon entering Game Master. Likewise, while also not a "setting" in the traditional sense, the state of the "Auto-Apply Strategies" button Game Master's "Scenario Injects" panel will also be remembered using the same logic, making it easier for OC/Ts to control when strategies should be manually applied. Notably, since these settings are saved to the browser rather than to GIFT's databases, they are shared by any GIFT instances running on the same machine as long as the same browser is used, which can be useful for developers. This also means that using a different browser or machine to interact with GIFT will require adjusting the settings again, as will deleting cookies for GIFT's websites. Overall, this change is aimed to speed up repeated visits to Game Master and avoid needing to reapply settings that are needed on a repeated basis.

Finally, Game Master has received several performance improvements aimed to both speed up user interface interactions and to fix a few bugs that could cause clients to lose their connections to the server. The largest of these improvements were made to Game Master's "Assessments" panel, which was found to be the largest bottleneck for webpage performance. Performance issues commonly came up during sections of a DKF session where assessments were fired in quick succession, typically from assessment conditions like the Muzzle Flagging condition that could repeatedly flip between Below Expectation and At Expectation assessments. The overall problem was that the "Assessments" panel would continually queue up rendering operations for every performance assessment that was received, even if the previously received assessment had not even been rendered yet. Generally, rendering operations only took a few milliseconds and were imperceptible to users, but performance assessments were sometimes coming in on the order of nanoseconds, causing hundreds of operations to be queued. If enough of these operations were queued, the browser could become unresponsive to user input, and if this went on for over 10 seconds, it would cause the server to think that the client had timed out and send it back to the login screen. This was fixed in three



## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

ways. First, if the “Assessments” panel is still in the middle of applying rendering operations, it no longer kicks off more rendering options when performance assessments are received. Instead, it saves the data from all of the performance assessments that are received in the meantime, and once the last set of rendering operations has finished, it performs a single, combined set of rendering operations to visually apply all of the performance assessments that were received, reducing the overall number of rendering operations. Secondly, each set of rendering operations is only handled after user input and server responses have been processed, which should ensure that the page remains responsive even during intense rendering. Thirdly, the timeout configuration settings on the server have been improved to reduce the likelihood of a client being logged out just because it is busy. With these changes combined, Game Master is overall more responsive, performs many rendering operations faster, and is less likely to encounter erroneous behavior when observing scenarios with rapid performance assessments. See Figure 11 for a comparison of rendering performance in the 2022-1 and 2023-1 GIFT releases.



**Figure 11.** This is a side-by-side comparison the rendering performance in the 2022-1 release on the left and the new 2023-1 release on the right. The same scenario is being viewed for AAR playback in both versions, but the 2023-1 release on the right has finished its rendering in a few seconds, while the 2022-1 release on the left is stuck rendering for over a minute and will eventually time out and show an error.

### Third Party Updates

Several of GIFT’s core Java library dependencies have been updated to more recent versions to keep up with current security requirements. These will not impact most end users in a noticeable way, but they will impact developers and system administrators. Of particular note, GIFT’s Log4J version has been updated to 2.20.0. This impacts nearly all of GIFT’s codebase, since Log4J, in combination with SLF4J, provides the APIs that are used to build most of GIFT’s developer logs. Fortunately, external developers using GIFT that have their own custom code should not need to update any API calls that they make to Log4J or SLF4J, since GIFT’s new dependencies include a converter to automatically handle legacy API calls.

Updating Log4J had knock-on effects for several of GIFT’s other dependencies. ActiveMQ was upgraded to version 5.18.3 to coincide with the new Log4J version, which also required updating GIFT’s ActiveMQ configurations to allow GIFT’s modules to continue messaging one another. These updates should only impact GIFT instances with custom ActiveMQ configurations, which are not particularly common outside of server instances like GIFT Cloud. The Guava and Hibernate libraries were also minorly impacted by the Log4J update, but both were addressed simply by removing or replacing a few supporting JARs that used Log4J APIs, so neither needed a full version upgrade.

Users downloading GIFT from [gifttutoring.org](http://gifttutoring.org) or developers creating new branches likely will not see a noticeable change in behavior from this update, but developers with existing branches from before this change will likely feel the impact the next time they attempt to merge GIFT’s Subversion (SVN) trunk into their branches. While most of GIFT’s source code received only minor changes that should merge cleanly, the updates to GIFT’s build configurations and scripts may cause SVN merge conflicts in branches that

have modified GIFT's class paths to include new libraries. Any developers that run into such merge conflicts or that would like assistance merging the latest from trunk are encouraged to reach out to [the troubleshooting forum on gifttutoring.org](http://the.troubleshootingforum.org).

## **REQUESTED FEATURES FROM GIFTSYM11**

---

GIFT is community-driven, and we take pride in our user base, especially as it relates to functions and processes requested to support their research and content delivery needs. From last year's symposium, there were relatively few papers which actively requested or demanded features for development. This is good and shows a robust platform – the majority of papers presented described an activity which is ongoing with GIFT, rather than addressing some weakness or shortfall.

## **GIFT AND IEEE STANDARDS ON ADAPTIVE INSTRUCTIONAL SYSTEMS**

---

The discussion continues on adaptive instructional systems through the IEEE Learning Technologies Standards Committee (LTSC). LTSC coordinates with other organizations that produce specifications and standards for learning technologies. The GIFT community invites the reader to join the conversation on what data exchange standards for learning technologies might look like in the future.

## **CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH**

---

The GIFT program has seen significant advancement since its conception in 2011. Each year, the community continues to develop new features and use cases for adaptive instructional systems. With a near-term focus on utilizing GIFT to address data use and team tutoring challenges, we are excited to continue evolving the tools and methods to address critical capability gaps to drive future training requirements and system development. Stay tuned for continued improvements that address all facets of intelligent tutoring in today's education and training climate. Check back next year to see what kind of progress we make!

## **ACKNOWLEDGEMENTS**

---

The research reported in this document was performed in connection with contract number W912CG-20-C-0021 with the U.S. Army Contracting Command – Aberdeen Proving Ground (ACC-APG). The views and conclusions contained in this document are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of ACC-APG, CCDC-SC STTC 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 herein.

## **REFERENCES**

---

Brawner K, Ososky S. (2015). The GIFT 2015 report card and the state of the project. In: Sottolare R, Sinatra A, editors. Proceedings of the 3rd Annual GIFT Users Symposium (GIFTSym3); 2015 Jun 17–18; Orlando, FL. Aberdeen Proving Ground (MD): Army Research Laboratory (US); c2015. ISBN 978-0- 9893923-8-9.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Brawner, K., Sinatra, A. M., & Sottolare, R. (2017). Motivation and research in architectural intelligent tutoring. *International Journal of Simulation and Process Modelling*, 12(3-4), 300-312.
- 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., Hoffman, M., Nye, B., & Meyer, C. (2019). *Architecture and Ontology in the Generalized Intelligent Framework for Tutoring: 2019 Update*. Paper presented at the Proceedings of the 7th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym7).
- Goldberg, B., Brawner, K., & Hoffman, M. (2020). *The GIFT Architecture and Features Update: 2020 Edition*. Paper presented at the Proceedings of the 8th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym8).
- Goldberg, B., Roberts N., & Lenz, T. (2022) *The GIFT Architecture and Features Update: 2023 Edition*. Paper presented at the Proceedings of the 10th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym10).
- Hoffman, M., Goldberg, B., & Brawner, K. (2021). *The GIFT Architecture and Features Update: 2021 Edition*. Paper presented at the Proceedings of the 9th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym9).
- Ososky, S., & Brawner, K. (2016). The GIFT 2016 community report. Paper presented at the Proceedings of the 4th Annual GIFT Users Symposium.
- Roberts, N., Lenz, T., & Goldberg, B. (2023, July). The GIFT Architectural and Features Update: 2023 Edition. In *Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11)* (p. 3). US Army Combat Capabilities Development Command–Soldier Center.
- Sottolare, R., Brawner, K., Goldberg, B. & Holden, H. (2012). The Generalized Intelligent Framework for Tutoring (GIFT). US Army Research Laboratory.

## ABOUT THE AUTHORS

---

*Nicholas Roberts is a senior software engineer at Dignitas Technologies and the engineering lead for the GIFT project. Nick has been involved in the engineering of GIFT and supported collaboration and research with the intelligent tutoring system (ITS) community for nearly 11 years. Nicholas contributes to the GIFT community by maintaining the GIFT portal ([www.GIFTTutoring.org](http://www.GIFTTutoring.org)) and GIFT Cloud ([cloud.gifttutoring.org](http://cloud.gifttutoring.org)), supporting conferences such as the GIFT Symposium, and participating in technical exchanges with Soldier Center and their contractors.*

*Benjamin Goldberg, Ph.D. is a senior research scientist at the U.S. Army Combat Capability Development Command – Soldier Center, and is co-creator of the Generalized Intelligent Framework for Tutoring (GIFT). Dr. Goldberg is the technical lead for a research program focused on the development and evaluation of Training Management Tools for future Army training systems. His research is focused on the application of intelligent tutoring and artificial intelligence techniques to build adaptive training programs that improve performance and accelerate mastery and readiness. Dr. Goldberg has researched adaptive instructional systems for the last 15 years and has been published across several high-impact proceedings. He holds a Ph.D. in Modeling & Simulation from the University of Central Florida.*



# Integrating Mixed Reality, Physical Simulators, and Adaptive Learning: A Use Case for Implementation

Michael Cambata<sup>1</sup>, Thomas Lenz<sup>1</sup>, and Randall Spain<sup>2</sup>

Dignitas Technologies, LLC<sup>1</sup>, U.S. Army Combat Capability Development Command (DEVCOM) – Soldier Center<sup>2</sup>

## INTRODUCTION

---

Low-cost, lightweight, integrated mixed reality (MR) solutions for individual and collective simulated training are becoming more prevalent in the modeling, simulation, and training (MS&T) landscape and will serve an integral role in meeting the requirements outlined by the Army Learning Model 2030-2040 (U.S. Army, 2024) and the U.S. Navy’s Ready Relevant Learning (RRL) initiative (Department of the Navy, 2022). By combining immersive virtual environments with tangible physical interactions, MR training solutions allow learners to engage in realistic scenarios that bridge the gap between procedural and tacit knowledge without the burden of safety, budget, or equipment availability concerns. Integrating these solutions with adaptive training capabilities offers significant promise for creating highly effective and engaging training experiences. In this paper, we discuss a recent effort that aimed to develop an immersive MR-based training solution with an open-sourced adaptive training architecture to support adaptive MR-based crew gunnery training. Our use case centers on an M1A2 Abrams MR-based training station with software centric virtual representations of crew stations in which crew perform Basic Gunnery Skills against specified Gunnery Tables. Trainee Tasks include target acquisition, target designation, and engagement of stationary and moving targets. We discuss how we leveraged the Generalized Intelligent Framework for Tutoring (GIFT) to guide automated and semi-automated assessment of taskwork and teamwork skills and to deliver feedback and coaching to trainees. We also discuss the tasks and challenges associated with creating the adaptive MR based training system.

## SYSTEM INTEGRATION OVERVIEW

---

During this effort, several pre-existing systems were modified to work with each other. See Figure 1, below for a system design and communication diagram. The primary components were the following:

- **Mixed Reality Tactical Trainer (MRTT):** a set of hardware modules designed to mimic realistic M1A2 Abrams crew stations (Gunner, Driver, and Commander). The hardware includes an interface for hardware inputs and manipulation.
- **Software-centric Immersive Virtual Environment (SIVE):** a lightweight, portable low-cost human interface for point-of-need collective training provides a mechanism for interfacing across multiple modalities: web pages, game controllers, virtual reality (VR), and mixed hardware via software.
- **Simulation Systems:** a simulation system backend provides the simulated exercise for trainees to perform. The Close Combat Tactical Trainer (CCTT) provides the core engine for simulating vehicle functionality of the M1A2. Instances of the CCTT OneSAF Semi-Automated Forces (SAF) and OneSAF Computer Generated Forces (CGF) hosts provide constructive forces for our virtual environment.
- **GIFT:** an intelligent tutoring system capable of connecting to external assessments, reading data during training, and automatically performing assessments.

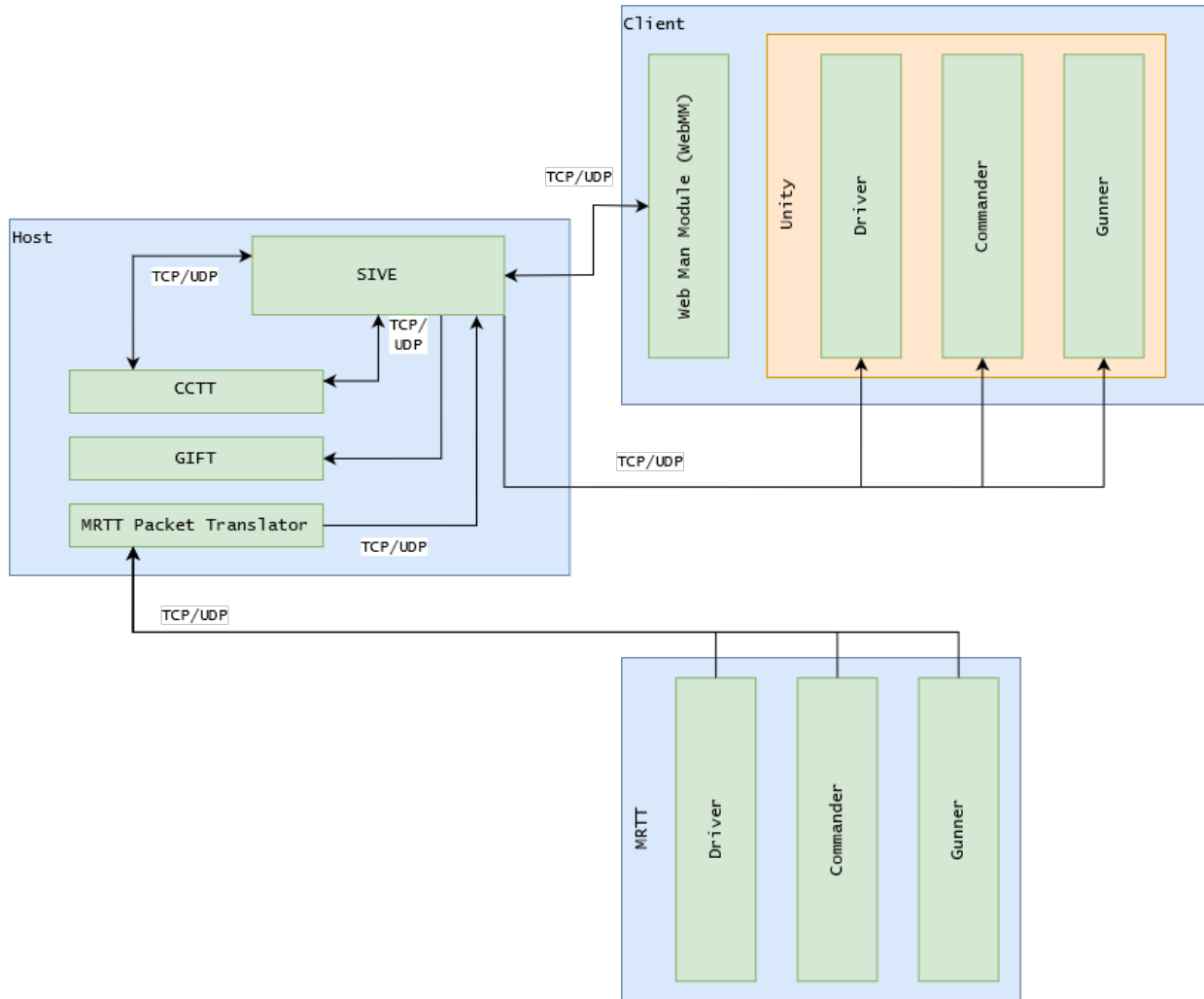


Figure 1. Mixed Reality Simulation System Component Diagram

## TRAINING USE CASES

The research's training use case focused on demonstrating how SIVE, MRTT, and GIFT contribute to the automated assessment of crew gunnery skills. For this purpose, we developed a virtual crew gunnery scenario, requiring crew members to engage both stationary and moving targets using the vehicle's integrated weapon system. Trainees assume designated crew roles—Vehicle Commander, Driver, or Gunner—and complete the scenario at their respective MRTT stations. Refer to Table 1 for a listing of the specific assets used for the discussed use case. In the table, “Yes” indicates that item was used for this research, “Available” indicates that it is interchangeable with the configuration, and “Unused” indicates that modality does not apply to that role.

The commander is responsible for managing the tank's operations and directs the crew. This commander station is equipped with multiple periscopes and a joystick-controlled thermal night vision viewer; the commander surveys the environment and monitors the tank's systems and location via an integrated display. The gunner is responsible for engaging enemy targets by operating the vehicles weapon system and laser range finder which facilitates precise distance measurement. Additionally, the gunner operates the front

machine gun and checks the main gun's status. The loader selects ammunition from the storage and loads it into the main gun, following the gunner's instructions on the type of round required. The driver steers the tank. Each of these positions is represented in the corresponding MRTT pod station except for the loader position, which is simulated using a low fidelity platform that will be described in more detail later. Trainees view the virtual environment and their corresponding station controls using a VR (virtual reality) or AR (augmented reality) head-mounted display (HMD) or through a monitor.

**Table 1. Training Use Case Equipment to M1A2 Crew Position Usage**

Role	Gaming PC	Vive Pro 2 HMD	MRTT Physical	WebMM	Xbox Controller
Driver	Yes	Available	Available	Yes	Yes
Gunner	Yes	Yes	Yes	Available	Unused
Vehicle Commander	Yes	Yes	Yes	Yes	Unused
Loader	Yes	Unused	Unused	Yes	Unused

The scenario required the crew to navigate the tank along a predefined route and actively work together to scan, detect, acquire, and engage targets (Figure 2). Once a target is acquired the crew must quickly engage it using the correct weapon type and corresponding ammunition while completing the steps outlined in the direct fire engagement process. A critical component of the task is coordinating search sectors among crew members, swiftly discerning target types, and selecting suitable ammunition and weapon system to designate and engage the target based on its range and lethality.



**Figure 2: Visualization of route and target locations.**

Table 2 lists the targets included in the scenario as well as the ideal weapon system and ammunition type that should be used to engage opposing forces based on the range, or distance away from the trainees. Using this information, GIFT can be used to assess whether crews selected and used the appropriate weapon system and ammunition.

**Table 2. Engagement Targets and Corresponding Weapon System and Ammunition Type**

Engagement	Target Type	Gun	Ammo	Range
1	Truck	COAX	.50 cal	Short
2	Truck	COAX	.50 cal	Short
3	Tank	Main	SABOT or HEAT	Short
4	Tank	Main	SABOT or HEAT	Long
5	Tank Truck	Main COAX	SABOT .50 cal	Long

### Assessing Crew Duties

GIFT includes a feature for creating and implementing real-time assessment models to facilitate automated and semi-automated evaluations of tasks in simulation-based training. This is accomplished by generating a Domain Knowledge File (DKF) through GIFT's course authoring tool. This process involves decomposing a training domain into a series of concepts, tasks, and subtasks and designating condition logic embedded in GIFT to facilitate task assessment. The goal of this research was to assess individual crew duties and team-level coordination skills, which are critical to crew performance. Crew coordination skills encompass principles, attitudes, procedures, and techniques that turn individuals into an effective team. These skills include:

- Communication: Ensuring clear, concise verbal exchanges, active listening, and standardized phraseology among crew members.
- Leadership and Followership: Enabling crew members to lead or follow as needed, based on the mission's context, promoting mutual respect and cooperation.
- Situational Awareness: Keeping a precise awareness of the environment, mission goals, and other pertinent details for informed decision-making.
- Decision-Making: Making decisions that are timely and suitable, taking into account safety, mission demands, and resource constraints.
- Task Management: Prioritizing tasks and responsibilities effectively, aligned with mission needs and crew abilities.
- Teamwork: Collaborating to achieve shared objectives, nurturing a culture of trust, respect, and mutual aid.

To support our demonstration, we developed a DKF to evaluate crew performance outcomes and coordination during engagements, detailing each crew member's roles and responsibilities. To guide this process, we outlined the roles and responsibilities of each crew member during an engagement (Table 2) and aligned these activities with condition class features in GIFT's DKF structure, noting which condition class logic could be used to support assessment and if new condition class logic needed to be developed to support our assessment goals.



Table 3. Crew Member Roles and Responsibilities

Phase	VC	GNR	LOADER	DRIVER
<b>Target Acquisition Process:</b> Crew Search	Search for targets	Search for targets		
<b>Target Acquisition Process:</b> Target ID and Engagement Decision	Engagement decision			
<b>Engagement Sequence:</b> Fire Command	<ul style="list-style-type: none"> <li>Alerts Firer (Announces GNR)</li> <li>Announces Ammo</li> <li>Announces Target</li> <li>Announces Direction*</li> <li>Prioritizes Targets(s)</li> </ul>	<ul style="list-style-type: none"> <li>Confirm fire control switch to normal</li> <li>Confirm LRF to ARM 1st RTN</li> <li>Sets weapon selection switch to appropriate gun</li> <li>Sets ammo selection switch to appropriate ammo</li> <li>Acquires target through sights (3x mag)</li> <li>Announces <b>IDENTIFIED</b></li> <li>Centers target in GPS</li> <li>Switches to 10x mag</li> <li>Orients weapon towards target</li> <li>Lases target (Ranging action) and evaluates range return</li> </ul>	<ul style="list-style-type: none"> <li>Ensures GUN/TURRET DRIVE is in POWERED position</li> <li>Ensures Yellow ARMED light is on</li> <li>Loads Ammo</li> <li>Announces UP</li> </ul>	<ul style="list-style-type: none"> <li>Observes</li> </ul>
<b>Engagement Sequence:</b> Direct Fire Engagement	<ul style="list-style-type: none"> <li>Announces EXECUTION command</li> <li>Observes</li> <li>Sensing term (TARGET, DOUBTFUL (LEFT or RIGHT), LOST, OVER, SHORT)</li> </ul>	<ul style="list-style-type: none"> <li>Announces ON THE WAY</li> <li>Waits one second and fires</li> <li>Sensing term (TARGET, DOUBTFUL (LEFT or RIGHT), LOST, OVER, SHORT)</li> </ul>	<ul style="list-style-type: none"> <li>Remains clear of recoil path</li> <li>Observes</li> <li>Sensing term</li> </ul>	<ul style="list-style-type: none"> <li>Observes</li> </ul>

A primary objective of this research was to extend GIFT’s assessment capabilities to assess team competencies through MR training stations. As an example, we wanted to assess critical task work duties aligned to the gunner position such as designating the fire control switch, using the laser designator, setting the weapon selection switch to the appropriate gun prior to an engagement and setting the ammo switch to the appropriate ammunition and the ability to assess crew coordination competencies such as shared situation awareness. In addition, we wanted to demonstrate how we could use GIFT in concert with external

assessment tools such as Team Communication Analysis Toolkit (TCAT) to assess crews' information exchange and communication behaviors during the fire command component of the direct engagement process for each engagement. For this research, crew coordination was assessed manually within GIFT. The fire command sequence -- an important component of a crew's coordination activities - is a well-defined protocol for communicating information and actions between team members to coordinate the effective engagement of a target. TCAT is discussed in detail in our future integration and research conclusions section.

In the following sections, we discuss the system integration components required for demonstration followed by a discussion of challenges and lessons learned along the way.

## SYSTEM COMPONENT DETAILS

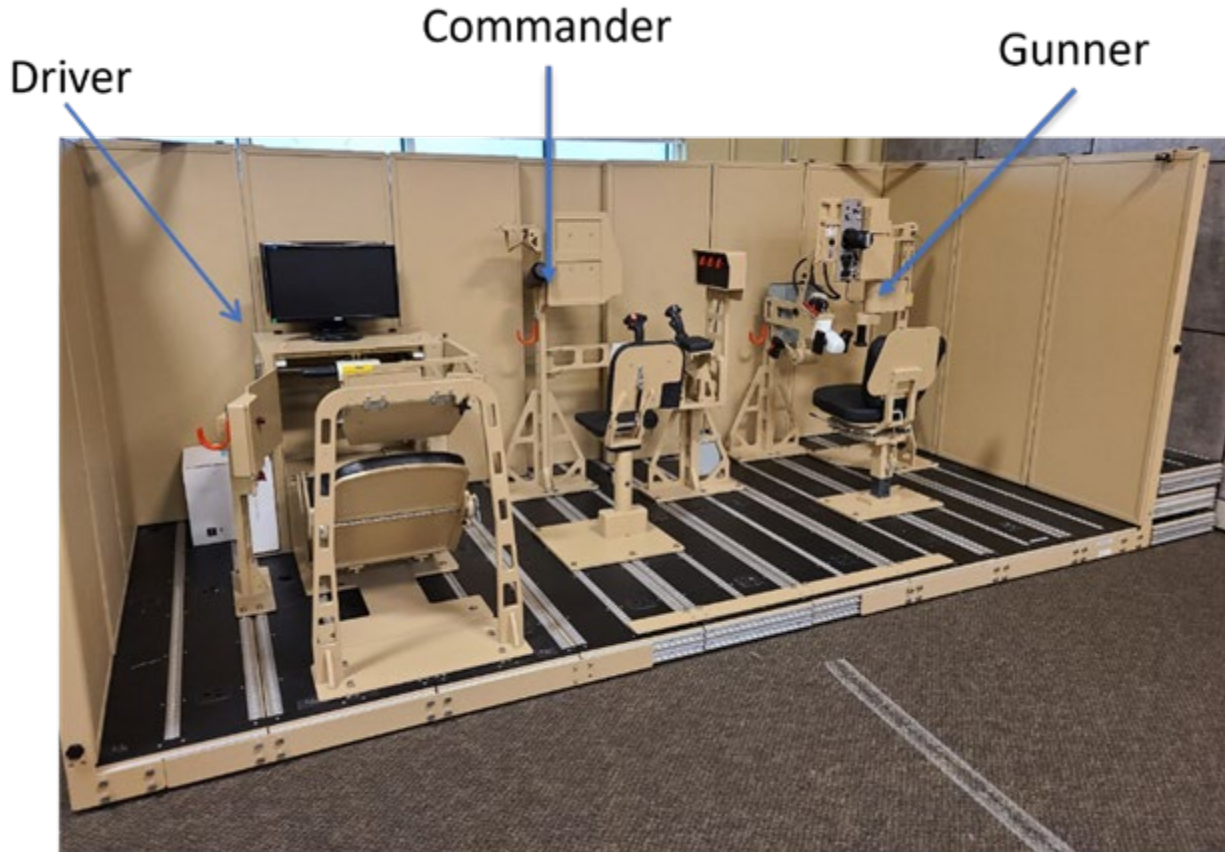
---

### MRTT

MRTT combines MR displays, user input and output technologies within a common hardware configuration to support the delivery of effective training at optimal fidelity. The trainee sits in what visually look like a realistic crew station and interacts with a combination of real or virtual controls and instruments enabled through the MR technology. The MR technology allows for visual representation of different levels of training guidance dependent on the user. The MRTT includes 3 hardware modules, or pods, associated with the crew stations for the M1A2 SEP V2 Main Battle Tank (MBT).

- Gunner
- Driver
- Commander

Each module consists of the physical hardware associated with the pod, and an interface computer that communicates the state of the physical hardware to the simulation via network data packets. For example, the driver pods send information regarding the brake, steering wheel position, and gear shift information. The MRTT hardware includes most of the physical controls required to operate a virtual instance of the M1A2. It was designed to be used alongside a visual 3D environment which shows the full interior of the vehicle and its surroundings. The crew position hardware and networking components were the only aspects of the MRTT that were leveraged during this research as they provided the realistic tactile interactions as well as the interface converting physical analog inputs into digital outputs that SIVE uses to update simulation states. Specific controls provided by MRTT include: Driver: Gear Selection, Throttle, Brake (Main and Parking), and Steering; Gunner: Fire Control Mode, Trigger, Main Gun, Elevation and Azimuth, and various hand controls; Commander: CROWS functions, CITV toggles, Elevation and Azimuth, and Weapon Control Modes. Figure 3 shows the MRTT configured as one device with all three position pods connected.



**Figure 3. MRTT configured as one device with all three crew position pods connected**

## **SIVE**

SIVE is a lightweight, software interface that provides a communication layer between physical hardware in an extended reality training paradigm and VR, AR, and MR training applications. SIVE provides an immersive, human in the loop training (HITL) environment that conforms to the U.S. Army's Synthetic Training Environment's (STE) mission (low cost, small footprint, readily available outside of complex installations) as a contrast to the large manned-simulation programs, such as CCTT. SIVE's output provided a mature starting point for translating hardware inputs between physical hardware and software simulation capabilities.

In addition to providing a software communication and translation layer between the physical hardware and the backend simulation suite, SIVE also provides the virtual environment displayed within commercial off the shelf (COTS) VR equipment. SIVE leverages Unity's real-time 3D development engine because it provides a robust set of features and relative ease of development.

For our use case, SIVE simulates the vehicle interior of the M1A2 and serves as the primary visual interface between the trainee and the training environment. The M1A2 interior is modeled with a high degree of physical fidelity – including accurate representations of the buttons, switches, levers, and indicators for each crew member position - to provide a truly immersive experience. As trainees interact with physical controls, the virtual representations update their state to align with those inputs. This provides the trainee with the necessary cues to understand the current state of the vehicle and its systems.



Figure 4. MRTT Physical Hardware Shown on the Left with Trainee VR HMD Display Shown on the Right

SIVE's flexible architecture makes it possible to dynamically mix and match levels of modalities dynamically, e.g. web page in use for the loader, but immersive for the commander, driver, and gunner. In fact, the architecture supports multiple instances of a given station being active at once, thus making it possible for an immersed operator to receive assistance from a facilitator using a web page. See Figure 4 for an example of both the physical hardware, and the VR HMD display. As a part of the SIVE environment, the "Web Manned Module" or WebMM, is hosted on a virtual machine (VM) for use by locally networked computers, and offers a virtual interface to control simulated crew positions. This simple interface provides a mechanism to interact with crew positions not physically provided. For example, the loader position was not built out with physical hardware but loader actions can be performed using this webpage and a mouse to select types of ammunition, open and close the breach, load rounds, and set the state of safety breach. The interface is shown in Figure 5.



Figure 5. SIVE's WebMM Loader position interface

## Simulation Systems

In past research and experimentation with mounted ground training instances, our team leveraged the use of Virtual Battle Space (VBS) due to its existing GIFT integrations. However, VBS lacks multiple vehicle crew position capabilities. As such, our research leveraged simulation back end instances of CCTT and OneSAF, running on VMs, to simulate the battlefield, vehicles, and weapons within the scenario.

The Army currently leverages monolithic manned-module simulators such as CCTT and Aviation Combined Arms Tactical Trainer (AVCATT) virtual and collective training to support collective training exercises. Both CCTT and AVCATT rely heavily on complex cabin physical hardware to recreate the crew and cockpit stations for training, which require a large foot-print during transport, use, and at idle. In addition to the form factor, these crew trainers rely heavily on instructors to provide performance feedback after action, with no automation.

By directly reusing the CCTT virtual and constructive baseline assets, we ensure that the fidelity of the simulated experience is on par with current Army training capabilities. As in-development STE capabilities mature, the simulation system will shift to support current state of the art training capabilities.

## GIFT

GIFT has been used for nearly a decade to design and investigate intelligent tutoring capabilities in immersive training environments. As an open-source modular framework, researchers have used GIFT to design intelligent tutoring-based training courses to teach marksmanship fundamentals (Goldberg & Amburn, 2015), land navigation and terrain association skills (Goldberg & Boyce, 2018), dismounted infantry battle drill fundamentals (Folsom-Kovarik & Sinatra, 2020), and counterinsurgency doctrine (Spain et al., 2022). In recent years, GIFT has been enhanced to support intelligent tutoring for teams (Sottolare et al., 2018). These enhancements include the addition of a team modeling team structure in GIFT's Domain Knowledge File (DKF), new condition classes and scenario adaptations to support team assessment, the ability to deliver feedback at the individual and team levels, an instructor dashboard, referred to as the GIFT Game Master Interface that facilitates a "human in the loop" for assessing performance and injecting scenario adaptations during collective simulation-based training events, and the ability to facilitate improved after action reviews (AARs) of team performance using multimodal-based assessment capabilities.

## Integration

The components discussed above were developed separately from this effort, but were integrated to create a new configuration for simulation-based training. In this configuration, the MRTT hardware is used, but its outputs are being sent to SIVE. The state of the vehicle and the simulated scenario are used to update the SIVE provided M1A2 virtual scene, and are also being sent to GIFT for assessment. Figure 1, above, outlines the communication protocols in place and the flow of information through the system.

While these existing components can be integrated to achieve this functionality, development was required in order to do so. The steps of that development are discussed below.

## ADDITIONAL DEVELOPMENT

---

In order to take these systems and integrate them, additional development was necessary. Some of this development was performed through the creation of new applications designed to read, translate, and

transmit the various messages to and from the necessary components. After that, modifications were made to GIFT to allow it to interface with those applications.

### **Standalone Applications**

Two Windows applications were developed, designed to run on the computer that hosts SIVE and allow the SIVE system to connect to both the MRTT equipment and GIFT.

The **MRTT Packet Translator** reads incoming data from the MRTT hardware, which represents changes in the states of simulated vehicle controls. These packets are translated to a format useable by SIVE, and then sent from the MRTT Packet Translator to the SIVE system.

The **SIVE Socket Server** allows SIVE to send messages to GIFT, in two formats. One format is specific to SIVE and represents the current state of any controls simulated in SIVE. The other is in the form of protocol data units (PDUs) in the Distributed Interactive System (DIS) format.

Initially, the socket server only contained support for SIVE's messaging format, under the assumption that these control states would be the most critical use of GIFT's interaction with SIVE. However, as the effort to integrate GIFT and SIVE continued, it became clear that GIFT should be able to perform assessments on the underlying simulation that SIVE is interfacing with, as well. Because that simulation sends DIS PDUs to communicate, the DIS support was added to the SIVE Socket Server.

### **GIFT Modifications**

The applications discussed above allow the necessary information to be sent between GIFT, SIVE, and the MRTT hardware. However, modifications to GIFT were necessary in order to properly use the data it receives from SIVE.

#### ***Gateway Module Plugin***

In order for GIFT to use data from SIVE, a new plugin was created for GIFT's Gateway Module to handle connections to the SIVE Socket Server. GIFT uses these plugins to connect to external training applications in other cases. The SIVE plugin manages the connection to the SIVE Socket Server and listens for DIS PDUs from that source. When incoming DIS data is received, the plugin triggers the appropriate response, including formatting the DIS data into a GIFT message that is sent to other modules within GIFT.

#### ***Support for Designator PDU***

A new GIFT condition was created for this plugin. The MRTT equipment and backend CCTT simulation support a laser rangefinder function for designating a target. This laser designator is associated with a DIS PDU, so that data can be read and assessed by GIFT in a condition. However, it was determined that JDIS, the library which GIFT uses to process incoming DIS messages, did not contain full support for reading the Designator PDU's data. To support that and allow the condition to be created, GIFT's DIS support had to be updated.

Prior to this, GIFT contained support for both JDIS and OpenDIS libraries, but exclusively used JDIS for reading input in its DIS interface. OpenDIS supports the Designator PDU, but converting GIFT to use OpenDIS for all PDU input would be a major task, and one that would necessitate extensive testing to ensure that all DIS-related functionality remained functional afterward.

The DIS interface was modified to check a PDU's type value when it is received. If the value matches that of a Designator PDU type, then OpenDIS is used to parse it. Otherwise, JDIS is used, as it was previous to this update.

OpenDIS does appear to have more support than JDIS, so this particular structure offers the option to use it going forward, or to transition from JDIS to OpenDIS over a longer period of time, rather than as a single major task.

### ***Designator Condition***

Once the support for Designator PDUs was added to GIFT, a Designator condition was created. The Designator condition allows for the following inputs:

- A list of learners being assessed
- A list of targets to check against
- How close a firing line needs to come to a target to trigger an assessment
- Whether or not the assessment takes time since designation into account. If so:
  - The time to trigger an Above Expectation assessment
  - The time to trigger an At Expectation assessment

Whenever a detonation occurs, if it occurs on a target entity *or* if it comes close enough to the target to trigger an assessment, the condition checks whether or not that target was the most recently designated entity. If time is being taken into account, it also checks when the entity was last designated.

If time is not being taken into account, then the condition assesses to At Expectation if the current target is the most recently designated target of the firing entity, and Below Expectation otherwise.

If time is being taken into account, then the condition assesses to Above Expectation, At Expectation, or Below Expectation depending on whether or not the last designation is within the time bands specified by the course creator.

### ***Ammunition Logging Functionality***

In addition to the new condition's assessment functionality, it was given the ability to track ammunition usage. Whenever a detonation occurs while the condition is active, the condition extracts the detonation's time stamp, firing entity, target entity, and ammunition type. If the data (aside from time stamp) is not the same as the previous entry, it adds that data entry to a list, and when the domain session ends, it records that list in the domain session log. While this process is not used for automatic assessment, it allows the ammunition data to be extracted and analyzed after the course ends.

The check against adding duplicate data entries is to keep the list from being filled with nearly identical information. A rapid-fire weapon can send out a significant volume of messages, making the log difficult to parse. By checking if anything has changed since the previous data entry, we can include only distinct data points. Time stamp is excluded from this comparison, because the time stamp between detonations will likely be different even if all other data is the same.

## CHALLENGES

---

### Complex Interaction Between Multiple Systems

Integrating multiple disparate systems presents technical challenges due to the diversity of the technologies, the communication protocols each uses, and the types of data that need to be leveraged. Three key challenges emerged during the primary integration of the various disparate systems: system familiarity, networking, and configuration.

At the outset of the research program, the team had one goal in mind: integrate adaptive learning technologies with an existing physical system to demonstrate a fully immersive and complete training capability. Several systems were not immediately familiar to the development team. The challenge was to learn about each system or identify resources that could provide subject matter expertise. MRTT was the system that the team had the least amount of experience with; however, our engineering team used tools, like WireShark, to analyze network traffic from the MRTT system. Once that data could be interpreted, our team worked to translate that data into a format, or response, that the other systems could leverage. In addition, CCTT also posed challenges to the team due to its sensitive nature and complex VM architecture: it requires running several discrete components that make up the SAFs, CGFs, and battlefield simulation. Understanding the core requirements for that capability, and whether CCTT was the ideal back-end simulation, were key moments for the success of the effort.

Once the team understood the core capabilities and how each operated, we then set out to detail the network configuration. Two physical switches connected and powered the MRTT Arduino-based hardware and a third connected the GIFT host and clients. With all connections physically made, our team set out to properly configure the virtual local area network (VLAN); the CCTT simulation back-end runs via a series of VMs running and the network connectivity between each must be configured to point to specific IP addresses. As one example, the host computer requires an extra IP address assigned for each MRTT station being used, and they need to be specific addresses, because the equipment sends data to predetermined addresses. If any of these addresses are reverted to a prior configuration or a different machine acts as the host, all affected stations will also require modifications to their IP addresses, as well as configuration files referencing them.

### Authoring Complexity

Creating courses in GIFT that correspond to equivalent scenarios in training applications is a problem that GIFT has solved before. We have extensive support for systems such as VBS 3 which allow us to both receive data from them and send injections to the simulations in response to GIFT's assessments. That makes it possible to change the simulated scenario as certain conditions are passed or failed.

Because GIFT can read incoming DIS messages, we have support for receiving data from the simulation run by OneSAF and CCTT. However, we do not currently have support for injecting data from GIFT back into OneSAF simulations. This can be added in the future, but it is beyond the scope of this current effort.

In addition to this, while the engineers involved have some experience with OneSAF, it is limited compared to other simulation systems. We were able to overcome this by contacting and learning from other Dignitas employees with more OneSAF experience. Ultimately, we were able to create a scenario that met the needs of our research, along with a GIFT Course that corresponds to the scenario. While GIFT still lacks support for sending data to OneSAF, we were able to set up and trigger events within OneSAF to allow OPFOR entities to move to preset locations, achieving much of the same functionality that we would if we were



using environment adaptations from GIFT. Improvements can be made to lower the complexity of this authoring process, but this challenge did not prevent us from creating the scenarios we needed.

## **Management of Resources**

Because of the scope of the overall training system, physical resource management proved to be a key consideration for demonstration and use for research purposes. Research using GIFT has historically, and typically, only required laptops, mobile devices, and other equipment such as GPS and other small form factor sensors. In the case of the MR M1A2 simulation research, our team has had to consider the logistics required to transport, configure, test, and operate systems with comparatively more hardware.

In order to operate the system, the following equipment is required:

- MRTT equipment
  - Three crew position pods
  - Three Switches
  - Assorted Networking Cables
- Four High-End Gaming laptops
- Four USB-to-ethernet adaptors
- Two HTC Vive Pro 2 HMD
- Two display port-to-mini display port adaptors
- One Xbox controller

The MRTT system can be operated with less than three crew positions; however, to ensure a full experience, the remaining crew positions must be operated from a laptop. For a recent conference, we demonstrated the gunner position with the physical hardware and VR headsets but performed driving tasks using an Xbox controller and loader tasks via a web page. This allowed conference attendees to experience the system at it's fullest without having to transport or configure all pieces of the system.

## **Documentation**

Documentation of existing systems was available but not comprehensive; many of the networking requirements had to be discovered through trial and error or best guess. However, as we matured the integration, we expanded the documentation including extensive setup and troubleshooting methods. One regular error that we experienced early on was with networking configurations; we discovered a best practice to re-use the same machine for hosting sessions to reduce the changes required to networking, network adapter, and specific application configurations. In early work, our team struggled to consistently operate the systems but developed documentation with step-by-step physical setup, startup, and configuration sequence. One advantage of our team structure is our use of a separate team for testing. The engineering team, responsible for documentation, hands off the instructions to the test team for verification. These individuals provide a fresh and unbiased point of view for the quality and accuracy of documentation. We were able to iterate through the configuration and operation procedures several times and have reduced the time for setup from several hours to under half an hour; this comprehensive documentation also reduced the expertise requirements so that even a team unfamiliar with the disparate components would be capable of performing a physical configuration and startup of the system with little or no engineering support.

## CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

---

The results of this effort provided a set of software useable for training, as well as research and development. It also provided demonstrations of GIFT and the MRTT hardware. However, further improvements are possible.

### Team Communication Analysis Toolkit (TCAT)

One planned improvement is to use a system of speech-to-text software called Team Communication Analysis Toolkit (TCAT), which is currently being integrated with GIFT, in a course that supports an MRTT training scenario. TCAT is a natural language processing (NLP)-driven framework that leverages deep learning-driven NLP to automatically analyze team communication data, categorize the data using dialogue classification schemes, and provide summary statistics of critical team communication features that can be used to investigate team performance (Pande et al. 2023; Paul et al., 2023).

TCAT's framework provides several core services to support near-real time team communication analytics during training events, including speech to text translation, dialogue act classification, and a data visualization feature. It uses the Microsoft Azure speech-to-text cloud service to automatically produce a transcript of team members' spoken communication captured during a training event. The transcript is passed to TCAT's NLP framework that analyzes the dialogue and automatically assigns labels to each utterance to indicate dialogue act, the intent of the speech (e.g., provide information, acknowledgement, command), and information flow, how information was passed between different levels of the team's hierarchy (e.g., provide information down when the squad or team leader shares information with team members, request information up when team members ask for information from the squad or team leader). This labeled transcript is available to TCAT's data analysis interface, which computes frequency statistics for each labeled variable and provides a visualization interface that transforms these statistics into informative graphics (e.g., charts of the frequency of each label per speaker). The driving objective of TCAT is to integrate NLP-driven insights about team communication into the GIFT. This integration will enable more robust adaptive coaching, scaffolding, and assessment in GIFT.

Currently, the performance of trainees can be assessed as their actions cause the vehicle to interact with the simulated world, but their verbal communication is assessed manually. With the integration of TCAT, though, GIFT provides the capability to evaluate their communication skills during a scenario.

### General Improvements

We are also interested in expanding the SIVE plugin to enable additional GIFT conditions. Other uses of the Designation PDU are possible, but in addition to that, SIVE has the capability to send the states of its controls directly to GIFT as they are changed. By building upon this it may be possible to assess that the vehicle is operated according to a set procedure.

Another area for improvement is documentation. While the documents are useable to start up and run the software under ordinary circumstances, there have been scenarios where issues have occurred and the supporting documentation was not sufficient for troubleshooting. In some of these cases, an engineer familiar with the project has needed to investigate to determine the solution. While we have been documenting these issues among our own team, we can improve our general instructions by adding more details about troubleshooting procedures and the reasoning behind the configurations. That way, if something fails, we would be able to reference a document to find a cause even if the exact circumstances have not occurred before.

Similarly, while we mentioned earlier in this paper that we reduced our startup time to around half an hour, we do still require a number of steps to be performed manually. It would be beneficial to script startup actions, to reduce the chances of user error or confusion. While this system runs on several networked machines and it is unlikely for a single “Start” button to be feasible without a major effort, it might be feasible to reduce the number of manual steps to a manageable set.

It is worth noting, as well, that these recommendations apply to more than just our effort integrating MRTT, SIVE, and GIFT. We can use these lessons learned to streamline work on future systems as well. This provides an example of GIFT being used to support a mixed reality simulation. This experience as well as the knowledge that came from it may be applicable on any number of cases where GIFT is required to interface with multiple connected systems.

## REFERENCES

---

- Department of the Navy. (2022). MyNavyHR: Ready Relevant Learning. From <https://www.netc.navy.mil/RRL>
- Folsom-Kovarik, J. T., & Sinatra, A. M. (2020, May). Automating assessment and feedback for teamwork to operationalize team functional resilience. In *Proceedings of the 8th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym8)* (p. 126-134). US Army Combat Capabilities Development Command–Soldier Center.
- Goldberg, B., & Amburn, C. (2015, August). The application of GIFT in a psychomotor domain of instruction: a marksmanship use case. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym3)* (pp. 115-124). US Army Combat Capabilities Development Command–Soldier Center.
- Goldberg, B., & Boyce, M. (2018). Experiential Intelligent Tutoring: Using the Environment to Contextualize the Didactic. In *Augmented Cognition: Users and Contexts: 12th International Conference, AC 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part II* (pp. 192-204). Springer International Publishing.
- Pande, J., Paul, S., Min, W., Spain, R., & Lester, J. C. (2023, May). Improving Dialogue Classification Models to Support Team Communication Analytics in GIFT. In A. M. Sinatra (Ed.). *Proceedings of the Eleventh Annual Gift Users Symposium (GIFTSym11)* (pp. 127–136). US Army Combat Capabilities Development Command–Soldier Center.
- Paul, S., Spain, R., Min, W., Pande, J., & Lester, J. (2023, September). Evaluating the Classification Performance of Natural Language Processing-Driven Team Communication Analysis Models. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 67, No. 1, pp. 2181-2186). Sage CA: Los Angeles, CA: SAGE Publications.
- Sottolare, R. A., Shawn Burke, C., Salas, E., Sinatra, A. M., Johnston, J. H., & Gilbert, S. B. (2018). Designing adaptive instruction for teams: A meta-analysis. *International Journal of Artificial Intelligence in Education*, 28, 225-264.
- Spain, R., Rowe, J., Smith, A., Goldberg, B., Pokorny, R., Mott, B., & Lester, J. (2022). A reinforcement learning approach to adaptive remediation in online training. *The Journal of Defense Modeling and Simulation*, 19(2), 173-193.
- U.S. Army. (2024). The Army Learning Concept for 2030-2040 [PDF]. Retrieved from <https://adminpubs.tradoc.army.mil/pamphlets/TP525-8-2.pdf>

## ACKNOWLEDGEMENTS

---

The research reported in this document was performed in connection with contract number W912CG-20-C-0021 with the U.S. Army Contracting Command – Aberdeen Proving Ground (ACC-APG). The views

and conclusions contained in this document are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of ACC-APG, CCDC-SC STTC 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 herein.

## ABOUT THE AUTHORS

---

*Michael Cambata* is a senior software engineer at Dignitas Technologies. Michael has worked at Dignitas Technologies since 2016 and has been contributing to the GIFT baseline since 2020. In 2022 and 2023, he acted as the engineering lead for FLATT. He has also led the development of GIFT's integration with external applications, and supported research data collection events that use the GIFT software. Prior to his work on GIFT, he has acted as a software engineer on training and simulation projects involving MRTS 3D®, the F-35 Training Management System (TMS), and OneSAF.

*Thomas Lenz* is a senior systems engineer and project manager at Dignitas Technologies. Thomas leads analysis, design, and implementation of immersive training solutions; he also oversees research and development of new technologies that can be applied to legacy training challenges. He has been involved in various aspects of GIFT's development and testing for several years. His background includes over 10 years of instructional systems design for live and virtual training systems and systems engineering for Army and Navy use cases.

*Randall Spain, Ph.D.* is a Research Scientist in the Training and Simulation Division at the U.S. Army DEVCOM, Soldier Center. He holds a PhD in Human Factors Psychology from Old Dominion University. His research focuses on designing and investigating adaptive and intelligent training systems.



# **THEME II: COMPETENCY FRAMEWORKS**



# Developing a Squad Competency Framework in STEEL-R

Grace Teo<sup>1</sup>, Michael King<sup>1</sup>, Jennifer Solberg<sup>1</sup>, Benjamin Goldberg<sup>2</sup>, Gregory Goodwin<sup>2</sup>, Meghan O'Donovan<sup>2</sup>, and Clifford Hancock<sup>2</sup>

Quantum Improvements Consulting<sup>1</sup>, U.S. Army Combat Capabilities Development Command (DEVCOM) - Soldier Center<sup>2</sup>

## INTRODUCTION

---

Consider these two questions: (1) What are the characteristics of good and poor performance in a battle drill? and (2) What makes a Soldier or squad perform well (or poorly) in a battle drill? While both are important, the first question can be answered by examining the immediate outcomes of that battle drill or even by a thorough task analysis of Army handbooks and official doctrine. The answer is invariably specific to the battle drill of interest and does not provide much information about different variations of the battle drill, or different types of training exercises. For instance, favorable performance outcomes of an iteration of Battle Drill 2A: React to Contact (BD2A) conducted during the daytime may differ from that of the same drill executed at nighttime (Teo et al., 2022).

In contrast, answering the second question can provide insight into the Soldier's or squad's competencies that drive their performance. These competencies comprise trainable attributes such as the knowledge, skills, abilities and other attributes (KSAOs) that the Soldier draws upon to perform the tasks and functions of their position or role (Horey & Fallesen, 2004; Shavelson, 2010). Rather than simply describing the desired performance outcomes on a particular drill, addressing the second question results in understanding the Soldier or squad's KSAOs, which can be used to predict how they would fare on other types of battle drills. Efforts to track and assess Soldiers' competencies in addition to their performance outcomes align with a Soldier-oriented approach as it helps the Army better understand its Soldiers.

## Competency Frameworks

The competencies associated with a position are typically organized into a competency framework that depicts the relationships among them. More fleshed-out frameworks may include details about how the competencies can be measured and behavioral indicators for various proficiency levels on each competency (Goldberg, 2023). These additional details allow the competency framework to serve as a roadmap and provide a shared reference for the Soldier's development and training (Horey & Fallesen, 2004). For the Soldiers themselves, assessments based on the competency framework will show them the KSAOs in their areas of growth and their observable behavioral evidence. For their Commander or superiors, competency assessments reveal if the Soldier is ready to face the myriad of challenges expected in their position. For the training instructor, they inform the training needs of the Soldier. For the Army personnel officer, competency assessments help make personnel appraisals more transparent and facilitate career management discussions with the Soldier. For instance, if there is substantial overlap between the competency framework of a Soldier's current military occupational specialty (MOS) and that of another, the Soldier who is proficient in all competencies for their MOS may well be eligible for reclassification.

Broadly speaking, the competency framework within an MOS should apply across the tasks and drills that the Soldier in that MOS can be expected to perform even though the behaviors and measures indicative of those competencies may vary widely depending on the task or drill. For example, indicators of the *Leadership* competency in BD2A include behaviors such as a leader conveying flanking and assault plans to the squad. In contrast, in Battle Drill 6 Enter and Clear a Room (BD6), indicators of the same competency may entail a leader giving the command for when to enter a room to clear it. Even for the same drill, indicators of the same competency can differ according to the mode of the drill, such as whether it was

simulated in a virtual environment or executed in a live one. In a live iteration of BD2A, a squad leader (SL) conveying plans to their Alpha and Bravo teams typically involves them communicating by moving back and forth between the teams, which are usually some distance apart. This looks very different from an iteration of BD2A conducted virtually, which may merely require the SL to press a key or click their mouse to communicate between the teams and from a static location. The competency framework would still apply to both modes of the drill, even if the indicators differ.

## STEEL-R

One of the main aspirations of the Synthetic Training Environment Experiential Learning for Readiness (STEEL-R) program is to inform the Synthetic Training Environment (STE's) Training Management Tool (TMT). The scope of STEEL-R includes developing the infrastructure to support the processing and analyses of longitudinal data of Soldier-students collected from multiple training events and experiences conducted in different modes (i.e., live and virtual) over numerous occasions to build and update a persistent model of the Soldier-students' competencies throughout their military career (Goldberg et al., 2021). Such a Soldier-focused endeavor necessitates competency frameworks that specify the behavioral indicators of (i) the various competencies and (ii) the various proficiency levels of the competencies for both virtual and live versions of common drill tasks.

In this paper, we delve into the essential components required for a competency framework applied in STEEL-R and the ongoing efforts and future considerations involved in developing a competency framework for infantry squads in STEEL-R.

## COMPETENCY FRAMEWORK FOR STEEL-R

---

Components of a typical competency framework include (i) **the competencies** that are identified as being required for job success as well as the relationships among them that show which competencies are precursors to or enable other competencies (Robson et al., 2023), (ii) **the sub-competencies** which may contribute to the higher-order competencies, (iii) **the specific KSAOs** under the sub-competencies, and (i) **the measures or indicators of the competencies** that enable them to be tracked and assessed. The Hierarchical-Affective Behavioral Cognitive (H-ABC) model (Vatral et al., 2022), a working team competency framework for STEEL-R for infantry squads, illustrates these components. Level 5 comprises the measures and behavioral indicators of the team KSAOs in Level 4, which in turn contribute to the team sub-competencies in Level 3 that are subordinate to the team competencies in Level 2. Level 1 organizes the competencies in Level 2 into broad categories (see Figure 1). While the higher-order constructs tend to be more abstract, the lower-order sub-competencies tend to be more concrete and more observable from behaviors exhibited by the squad (Vatral et al., 2022).



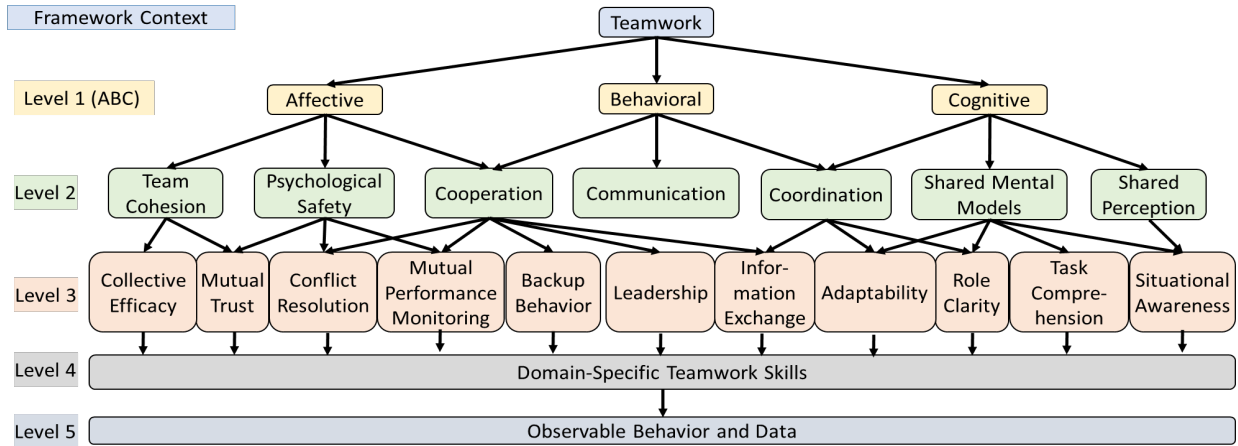


Figure 12. The H-ABC model (Vatral et al., 2022): a working team competency framework for STEEL-R for infantry squads.

### Small Unit Performance Analytics (SUPRA)

The Small Unit Performance Analytics (SUPRA) research program was instrumental to the competency framework development for STEEL-R as it contributed to the following: (i) establishment of the data pipeline, (ii) measure development and refinement, and (iii) competency review and validation.

#### *Establishment of data pipelines*

Under SUPRA, researchers leveraged advancements in sensor technology to collect various behavioral data that could yield evidence or indicators of competencies in the STEEL-R competency framework. Much data from the sensors was unavailable previously as they extended beyond the limits of human perception. For instance, data from the rifle-mounted inertial measurement units (IMU) detected and quantified the firing activity of each Soldier, and global positioning system (GPS) data pinpointed the location of each Soldier relative to the others and to each opposition force (OPFOR) throughout the drill. In processing and analyzing these data from the sensors, SUPRA researchers developed the data pipeline for ingesting and processing these live sensor data into STEEL-R’s Generalized Intelligent Framework for Tutoring (GIFT) so that sensor data can be used as competency evidence in STEEL-R assessments.

#### *Measure development and refinement*

Work in SUPRA involved developing measures and behavioral indicators from sensor-based data collected in live iterations of BD2A and BD6. Specifically, five key categories of sensor-based measures were derived from various data sources (see Table 1).

**Table 1. Five categories of measures in SUPRA**

<b>Data source</b>	<b>Description</b>	<b>Examples of measures</b>
1. Audio recordings of communications	Audio recordings from each Soldier captured verbal communications exchanged within squad	The proportion of procedural commands exhibited by the squad leader (SL), the number of firing commands from the Alpha team leader (ATL), the presence of <i>Shift Fire</i> callouts among the Alpha Team (ATeam), the number of words exchanged among the Bravo team (BTeam)
2. Inertial Measurement Unit (IMU) sensors	Rifle-mounted IMU sensors provide shot information from each Soldier. Helmet-mounted IMU sensors provide head orientation of each Soldier	<u>From rifle-mounted IMUs:</u> The firing rate of each Soldier, of the squad, of ATeam, and of BTeam. <u>From helmet-mounted IMUs:</u> Security coverage* by the ATeam, by the BTeam
3. Global Positioning System (GPS)	GPS coordinates provide location of each Soldier in time and space	Tightness of formation** of squad, of ATeam, of BTeam, Speed of movement of squad, of ATeam, of BTeam
4. Multiple Integrated Laser Engagement System (MILES)	Firing accuracy of simulated rounds from each Soldier	The number of Hits incurred by each Soldier, by ATeam, by BTeam, the probability of OPFOR Kills by each Soldier, by ATeam, by BTeam
5. Observer-Controller (OC)	Ratings on execution of task steps and on various constructs/competencies	<u>Task Step measures:</u> Ratings on how well “The squad leader develops a quick fire plan,” ratings on how well “The squad establishes local security” <u>Construct measures:</u> Ratings on squad’s Fire Effectiveness, ratings on the squad’s Information Exchange

\*Security coverage: defined by the degree to which the team’s Soldiers’ helmet orientations overlapped (a high degree of overlap would indicate that they were all facing the same direction and were not maintaining 360° security).

\*\*Tightness of formation: how closely the Soldiers in the squad or team are positioned to each other.

### ***Competency review and validation***

The H-ABC model (Vatral et al., 2022) was developed from an extensive review of teamwork literature. Separately, SUPRA researchers and Observer-Controllers (OCs) conducted a review of doctrine, field manuals, and Subject Matter Expert (SME) input to identify important constructs for infantry squads (O’Donovan et al., 2023). Since these constructs were applicable across infantry squad battle drills and described the KSAOs required to perform the drills, they are akin to competencies that the squads would draw upon to execute them (see Table 2). These OC competency constructs included an individual-level competency (i.e., weapons handling) and several competencies corresponding to constructs at different levels of the H-ABC model. For instance, in the H-ABC model, the domain-specific KSAOs in Level 4 would correspond to OC constructs such as fire effectiveness, cover and concealment, and control, which are essential for a team operating in the capacity of a squad in the domain of Army infantry, but not necessarily to teams from other occupation specialties. In contrast, there are OC construct competencies that are domain-agnostic and apply to teams in other occupation specialties. These correspond to the higher-level competencies (Levels 2 and 3) in the H-ABC model, and include the OC constructs of leadership, communication, information, supporting/backup behavior, and team orientation/cohesion. Incidentally, a review of command staff teams, i.e., a different MOS than the infantry squad, identified these same five Level 2 and 3 competencies as being important teamwork dimensions for command staff teams (Teo et al., 2021), providing a degree of validation for the H-ABC model of teamwork competencies.

**Table 2. Mapping the OC Competencies for infantry squads to the H-ABC model**

<b>SUPRA OC constructs</b>	<b>Construct definition from SUPRA</b>	<b>Level in the H-ABC model</b>
<b>Communication</b>	Mostly top-down use of formalized or expected channels of information	2
<b>Control</b>	Maintaining the ability to quickly maneuver or reconfigure combat power based on new or changing information, situations, and orders	4
<b>Cover and Concealment</b>	<u>Cover</u> : Protection from the effects of fires <u>Concealment</u> : Protection from observation or surveillance	4
<b>Fire Effectiveness</b>	Acquire the enemy’s location and mass the effects of direct fires to achieve decisive results in a close fight	4
<b>Information Exchange</b>	Mostly situational information conveyed through informal, bottom-up, or lateral lines of communication in support of mission accomplishment	3
<b>Initiative/ Leadership</b>	<u>Leadership</u> : Taking full responsibility for deciding and acting for the element to optimize mission accomplishment <u>Initiative</u> : Assuming responsibility, when appropriate, to ensure mission accomplishment	3
<b>Security</b>	Maintaining an instant ability to detect and proportionally react to tactical risks in any direction	4
<b>Simplicity</b>	Every decision and/or action has a decisive and necessary purpose	4
<b>Speed</b>	Accomplishing tasks as fast as possible without compromising the other tactical constructs (e.g., control, security, communication)	4
<b>Supporting Behavior</b>	Actions taken by individual Soldiers/leaders to provide support to other Soldiers or members of their group (tactical or otherwise)	3
<b>Surprise</b>	Striking the enemy at a time, place, or manner for which they are unprepared	4
<b>Violence of Action</b>	Combined use of speed, fire effectiveness, and control to maneuver combat power faster and more effectively than the enemy can counter	4
<b>Weapons Handling</b>	The safe and effective handling of, employment of, and firing of individual weapons during the conduct of military operations	Individual competency

## **COMPETENCY ASSESSMENTS**

---

In STEEL-R, data from a battle drill exercise are ingested and processed in GIFT and used to derive scores on the various measures. These scores are captured in xAPI statements and used by the Competency and Skills System (CaSS) for competency assessments. Hence, competency assessments can only be as good as the data and measures from which they are derived. Inferences and claims about competencies made from the scores pertain to the issue of validity, which must be addressed when interpreting and using these scores. Utilizing a working definition of competency as being the improvable KSAOs that an individual possesses and draws upon to perform a task under standardized conditions that resemble real-life situations, assessed against some benchmark of performance (Shavelson, 2010), the assessments should also specify the conditions of the task as well as the associated performance outcomes. However, any assessment event

from which data is collected is only a sample of possible assessment events, each with its sample of criterion, task-response pairs, occasion, rater, and assessment methods (Shavelson, 2010; Teo et al., 2023). This means that for competency assessments to be valid, it is critical that there are sufficient competency-relevant data collected under the appropriate conditions and that higher/lower scores on the assessment correspond to superior/poorer performance, respectively. The following sections address the concerns of task conditions or environmental factors and performance benchmarks.

### Task Conditions/Environment Factors

Currently, in STEEL-R, scores on performance measures (i.e., Level 5 of Figure 1) are categorized as being “Below Expectation,” “At Expectation,” and “Above Expectation” in GIFT (Robson et al., 2023). However, for the scores to be interpreted appropriately, contextual information about the tasking conditions must also be captured (Goldberg et al., 2021). For example, in a BD2A training event, the probability of hits on OPFOR by BLUFOR, P(Hits), is a measure of the squad’s shot accuracy, which reflects the squad’s understanding that an essential aspect of executing BD2A is suppressing and destroying the enemy (i.e., Task Comprehension in Level 3 of Figure 1). Suppose Squad A achieves a P(Hits) score of 0.5, and Squad B obtains a P(Hits) score of 0.2. Squad A may be performing “At Expectation,” while Squad B could be “Below Expectation.” However, if Squad B’s 0.2 P(Hits) score was attained in a night drill in foggy weather amidst thick vegetation and Squad A’s 0.5 P(Hits) score was from a day drill in the plains with good weather, then it could well be that Squad A was actually performing “Below Expectation,” and Squad B was performing “Above Expectation.” In this example, the interpretation of the P(Hits) scores was adjusted accordingly by adding relevant contextual information on illumination, visibility due to weather conditions, and vegetation density. Currently, STEEL-R only accommodates contextual information about whether the drill task was executed at night or during the day (Robson et al., 2023). To facilitate the appropriate interpretation of data and scores, it may be necessary to develop a taxonomy of common conditions for the various categories of measures (Robson et al., 2023). Table 3 could be a starting point for the common conditions for live drill tasks.

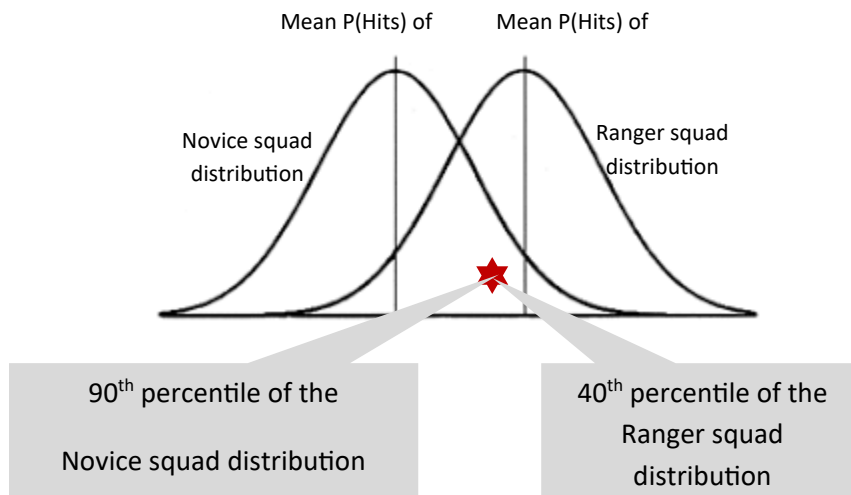
**Table 3. Task conditions that can impact the interpretation of measure scores in live drill tasks**

Category of measure	Examples of measures	Conditions to include for live drill tasks
1) Movement in time and space	Squad formation tightness, Squad speed of movement	<ul style="list-style-type: none"> <li>• Terrain flatness and altitude</li> <li>• Impeding terrain features (e.g., rivers)</li> <li>• Vegetation density</li> </ul>
2) Firing accuracy	P(OPFOR Hits), P(BLUFOR Hits), P(OPFOR Kills), P(BLUFOR Kills)	<ul style="list-style-type: none"> <li>• Time of day and season</li> <li>• Vegetation density</li> <li>• Weather conditions affecting visibility</li> </ul>
3) Communications	Proportion of procedural commands exhibited by squad leader (SL), <i>Shift Fire</i> callouts, <i>Last man</i> callouts	<ul style="list-style-type: none"> <li>• Environmental or ambient noises and sounds</li> </ul>

### Performance Benchmarks

Although crucial for meaningful competency assessments, there are currently no clear benchmarks for what constitutes good/moderate/poor performance on drill tasks. This is unsurprising since such benchmarks are challenging to define and would require task conditions to be specified. Nevertheless, one possibility is to leverage the performance of “known groups” under various conditions (Cizek, 2012). For example, if there are large volumes of data on how Ranger squads (i.e., known to be highly proficient infantry squads) and Novice squads (i.e., known to be low proficiency infantry squads) performed on BD2A that included the combination of the drill task conditions, then the respective distributions of performance scores can be generated. These distributions can then serve as two distinct norm groups against which the score of a new

squad (i.e., Squad C) obtained under the same drill task conditions can be compared (see Figure 2). Squad C's score (depicted by the red asterisk in Figure 2) would be at the 90<sup>th</sup> percentile of the Novice squad distribution but only at the 40<sup>th</sup> percentile of the Ranger squad distribution. Depicting the performance of different squad groups as distributions instead of absolute standards conveys the notion that there will always be variability in performance regardless of the squad's proficiency level. It is possible that Squad C could be a novice squad on a 'good' day or a Ranger squad on an 'off' day. This approach inevitably relies on large volumes of data, which is feasible given STEEL-R's data architecture. One advantage of this approach to defining performance with norm groups is that instead of labels such as "Below/At/Above Expectation" which may be somewhat challenging to envision, it describes performance in terms of actual groups of squads. Knowing how a particular squad performed relative to other squad groups facilitates relative or norm-referenced decisions, which decision-makers may have to resort to when criterion-referenced decisions are difficult due to challenges in defining absolute standards such as the standard for "At Expectation." In addition, by monitoring changes in the various norm groups over time, the Army would also gain a better understanding of each cohort of squads and Soldiers.



**Figure 13. Hypothetical distributions of P(Hits) measure scores from Novice and Ranger squad groups obtained from day drills with good weather and amidst dense vegetation.**

## CONCLUSION AND FUTURE DIRECTIONS

This paper covers considerations related to developing a squad competency framework to guide competency assessments in STEEL-R. Such a competency framework can set the foundation for training management strategies that are aligned with the modernization of training and the Army training aids, devices, simulators, and simulations (TADSS). We drew upon the work of various STEEL-R researchers and reviewed how research accomplished under the SUPRA work package contributed to the competency framework for STEEL-R. We discussed critical concepts in competency assessments, such as the validity of scores and performance benchmarks.

The next focus areas of work with the STEEL-R competency framework involve building on the SUPRA effort to specify and validate the alignment of KSAOs and behavioral indicators for an infantry squad executing BD2A, i.e., Levels 4 and 5 in the H-ABC model (see Figure 1), to the competencies in Levels 2 and 3. In addition, to accommodate data from training activities conducted in various modes, we will articulate the behavioral indicators and measures for virtual and live iterations of BD2A. We will develop approaches to incorporate various psychometric concepts into the longitudinal assessments of competencies

as well as examine the impact of using different algorithms for rolling-up scores between competency levels. To encourage the use of the competency framework as a team development roadmap, we will design data visualizations and performance feedback based on the competency assessments that support squad after action reviews (AARs) and selection of training activities. All these will inform specifications of the software requirements, including that for the xAPI data capture methodology, for STEEL-R software developer-partners. Studies will be conducted to produce data from virtual and live iterations of BD2A with which the STEEL-R pipeline (i.e., from drill execution to competency assessments to tools and visualization to facilitate squad feedback) can be validated. Analysis of this data can contribute to our understanding of the efficacy of different training activities for squads at various levels of proficiency.

Future work with the STEEL-R team competency framework should include the application of the framework to teams in other occupation specialties and an extension of the framework to encompass individual Soldier competencies. All these initiatives will ensure that the STEEL-R program will be well-positioned to contribute to the modernization of Army training and the STE.

## ACKNOWLEDGEMENTS

---

The research reported in this document was performed in connection with contract number W912CG-24-C-0003 with the U.S. Army. The views presented in this paper are those of the authors and should not be interpreted as presenting official positions, either expressed or implied, of the U.S. Government.

## REFERENCES

---

- Cizek, G. J. (2012). *Setting Performance Standards: Foundations, Methods, and Innovations*. Routledge.
- Goldberg, B. (2023). Drill-Practice-Repeat: Experiential Scaffolds. Workshop on Artificial Intelligence in Support of Guided Experiential Learning]. *International Conference on Artificial Intelligence in Education (AIED)*, Tokyo, Japan. [https://ceur-ws.org/Vol-3484/AIED-GEL23\\_paper\\_9\\_CEUR.pdf](https://ceur-ws.org/Vol-3484/AIED-GEL23_paper_9_CEUR.pdf)
- Goldberg, B., Owens, K., Gupton, K., & Hellman, K. (2021). Forging Competency and Proficiency through the Synthetic Training Environment with an Experiential Learning for Readiness Strategy. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (IITSEC 2021)*.
- Horey, J. D., & Fallesen, J. J. (2004). Leadership competencies for contemporary Army operations: Development, review, and validations. *46th Annual International Military Testing Association. Symposium Conducted in Brussels, Belgium*. <https://www.academia.edu/download/32824557/2004045P.pdf>
- O'Donovan, M. P., Hancock, C. L., Coyne, M. E., Racicot, K., & Goodwin, G. A. (2023). *Assessing the impact of dismounted infantry small unit proficiency on quantitative measures of collective military performance Part 1: Recommended test methodologies*. (Natick/TR-23/013). <https://apps.dtic.mil/sti/trecms/pdf/AD1210391.pdf>
- Robson, R., Corporation, E., & Goldberg, B. (2023). Digitizing Performance and Competencies. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (IITSEC 2023)*.
- Shavelson, R. J. (2010). On the measurement of competency. *Empirical Research in Vocational Education and Training*, 2(1), 41–63. <https://doi.org/10.1007/BF03546488>
- Teo, G., Jensen, R., San Mateo, C. A., Johnston, J., DeFalco, J., & Goodwin, G. (2021). Measures for Assessing Command Staff Team Performance in Wargaming Training. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (IITSEC 2021)*.
- Teo, G., Sikorski, E., King, M., & Solberg, J. (2023). Measurement Error and the Generalizability of Competency Assessments. *Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11)*.

Teo, G., Sikorski, E., Schreck, J., & Goodwin, G. (2022). Like day and night: Comparing squad level communications and shooting performance under differing battle drill conditions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66(1), 611–615. <https://doi.org/10.1177/1071181322661069>

Vatral, C., Biswas, G., & Goldberg, B. S. (2022). Multimodal Learning Analytics Using Hierarchical Models for Analyzing Team Performance. *Proceedings of the 2022 International Conference on Computer-Supported Collaborative Learning (CSCL), Japan*.

## ABOUT THE AUTHORS

---

**Grace Teo, Ph.D.**, is a Senior Research Psychologist at Quantum Improvements Consulting. Grace's research involves understanding and improving human performance under various conditions and in different contexts, such as working with different technologies, and in teams. Other research interests include assessments, decision-making processes and measures, vigilance performance, human-robot teaming, automation, and individual differences. Grace earned her Ph.D. and M.A. in Applied Experimental and Human Factors Psychology from the University of Central Florida.

**Michael King, Ph.D.**, is a Research Psychologist II at Quantum Improvements Consulting (QIC). While at QIC, Michael has led various training and human performance improvement projects, including a major initiative studying team communication and performance for the U.S. Army and evaluating virtual reality technologies for Air Force pilot training. Michael earned his Ph.D. in Experimental Psychology at Case Western Reserve University.

**Jennifer Solberg, Ph.D.**, is the CEO of Quantum Improvements Consulting, LLC. She has 15 years of military selection and training research experience, with an emphasis on leveraging innovative technologies for improving training in a measurably effective way. Her current research focuses on developing assessments of Warfighter performance to enable adaptive training, predictive modeling, and improved training effectiveness. She has led a team of senior government scientists in developing measures for identifying the cognitive and perceptual skills critical to visual IED detection. Jennifer earned her Ph.D. from the University of Georgia.

**Benjamin Goldberg, Ph.D.**, is a senior research scientist at the U.S. Army Combat Capability Development Command Soldier Center and is co-creator of the Generalized Intelligent Framework for Tutoring (GIFT). Dr. Goldberg is the technical team lead for a research program focused on the development and evaluation of Training Management Tool for future Army training systems. His research is focused on the application of intelligent tutoring and artificial intelligence techniques to build adaptive training programs that improve performance and accelerate proficiency and readiness. Dr. Goldberg has researched adaptive instructional systems for the last 12 years and has been published across several high-impact proceedings. He holds a Ph.D. in Modeling & Simulation from the University of Central Florida.

**Gregory Goodwin, Ph.D.**, is a senior research scientist with the U. S. Army Combat Capability Development Command Soldier Center at the Simulation and Training Technology Center (CCDC-SC-STTC), in Orlando, Florida. For the last decade, he has worked for the Army researching ways to improve training methods and technologies. He holds a Ph.D. in Psychology from Binghamton University and an M.A. in Psychology from Wake Forest University.





# Multimodal Measures for the Integration of Metacognitive Teamwork Processes During Simulation-Based Training

Megan Wiedbusch<sup>1</sup>, Ryan P. McMahan<sup>2</sup>, Anne M. Sinatra<sup>3</sup>, Benjamin Goldberg<sup>3</sup>,  
Lisa N. Townsend<sup>3</sup>, Joseph J. LaViola Jr.<sup>2</sup>, and Roger Azevedo<sup>1</sup>

School of Modeling, Simulation, and Training – University of Central Florida<sup>1</sup>, College of Engineering and Computer Science – University of Central Florida<sup>2</sup>, US Army Combat Capabilities Development Command (DEVCOM) – Soldier Center<sup>3</sup>

## INTRODUCTION

---

Teamwork is a critical and direct component driving the success of teams within extreme environments (e.g., military teams in war zones; Meslec et al., 2020). Teamwork includes a collection of cognitive, affective, verbal, and behavioral interactions between individual team members directed towards achieving a common goal (Kozlowski, 2018). The importance of teamwork for team performance has elicited the development of training methods and devices that aid in individual team members' ability to provide the skills necessary to effectively contribute to team performance (Vatral et al., 2022). For military training, this includes game-based learning environments (Martín-Hernández et al., 2021), wargames (Dorn et al., 2020), and live training (Johnston et al., 2022). In order for these training approaches to be effective, it is imperative that teamwork and team performance have valid and reliable measurements that can also be used to provide constructive feedback (Nonose et al., 2014). Typically, these measures are reliant predominately on behavioral markers that trained experts evaluate through observation. However, recently there has been a shift to introducing virtual simulations as effective training methods of teamwork for military scenarios (Balint et al., 2020; Johnston et al., 2019) as simulations offer several affordances that enhance how researchers are able to effectively observe, capture, track, and evaluate teamwork (real time and over time) for enhanced team performance outcomes (Goldberg et al., 2021). These new affordances may also provide a path by which less directly observable processes (i.e., metacognition underlying situational awareness) can be collected and integrated into teamwork metrics.

This paper proposes an extension of the hierarchical Affective, Behavioral, Cognitive (H-ABC) framework (Vatral et al. 2022) to incorporate metacognitive processes to further our assessment of teamwork within simulation-based training utilizing the Multimodal Observational OpenVR (MOOVR) Toolkit. We ground this expansion within a brief overview of our in-development, virtually simulated army battle drill (2A – Conducting a Squad Assault). Using this framework, we detail our approach to automatically assess individual-level and team-level performance using multimodal data that can then be integrated with GIFT (Generalized Intelligent Framework for Tutoring; Sottolare et al., 2012; Sottolare et al., 2017). Finally, this paper proposes several recommendations for potential metacognitive skills and competencies that can be operationalized with multimodal learning analytics derived from video, audio, gestures, head positioning, physiological responses (i.e., electrodermal activity and heart rate), log-files, and self-reports collected in (near) real-time during a simulation-based training exercise (Azevedo et al., 2018; Wiedbusch et al., 2023).

## METACOGNITION DURING TEAMWORK

---

Adaptation of one's cognition and behavior lie within metacognition, often colloquially defined as one's thinking about their thinking (Flavell, 1976; Winne & Azevedo, 2022). However, more specifically, metacognition refers to one's ability to (in)accurately reflect on, evaluate and control first-order cognitive processes (e.g., decision-making, perception, and memory; Katyal & Fleming, 2024). While a large body of research and theory on metacognition exists (Tarricone, 2011; Fleming, 2024; Norman et al., 2019), there

is still much debate over what does (and does not) constitute metacognition or a unifying framework that distinguishes between cognition clearly (Azevedo, 2020). However, across the many theoretical models of metacognition (e.g., Nelson & Narens, 1990; Winne, 2018) there are several metacognitive processes that can be roughly categorized as either monitoring/evaluative processes or as regulatory processes. Monitoring processes may include making evaluative judgements (e.g., feelings-of-knowing, judgements of learning, etc.) and reflection (Greene & Azevedo, 2009). Metacognitive regulatory processes may include selecting appropriate strategies, planning, and making changes to current learning/training approaches.

When performing as an individual, this monitoring and regulation allows us to adapt to volatile environmental factors (e.g., such as seen on a battlefield). The more accurate one is at making metacognitive judgements and evaluations, the more appropriately they can regulate their behavior, cognition, and affective processes which ultimately results in better performance outcomes (Fleming, 2024). When performing in a group, however, social metacognition, also referred to as team or group metacognition, expands these typical processes to include information processing and regulation about team performance, affect, and group dynamics (Folomeeva & Klimochkina, 2021; Thompson & Cohen, 2012). That is, in addition to monitoring and controlling our own knowledge, emotions, and actions, during social metacognitive processing, we are now additionally tasked with monitoring and regulating our team's knowledge, emotions, and actions. While many models of teamwork exist (e.g., Cooke et al., 2007; Endsley & Jones, 2001), many of these models represent cognition in teams as the sum of individual cognition while neglecting the cognitive factors that may influence cooperation (Nonose et al., 2014). As such, we see the opportunity to enrich these approaches with the affordances that virtual-reality (VR) simulations may provide to helping capture, measure, and provide feedback for less observable teamwork behaviors.

## **CASE-STUDY: BATTLEDRILL 2A – CONDUCTING A SQUAD ASSAULT**



**Figure 1. Example of the Battle Drill 2A drill being conducted in the UCF Arboretum (left) used to inform the design and development of the virtual-reality simulation environment (right).**

To contextualize the extension of this framework and the ongoing VR simulation development, we will use a case study of Battle Drill 2A as the task. In this drill, one squad leader and two infantry fire teams of four members each are moving as part of a platoon towards contact or an attack when the enemy initiates direct

fire. The squad's goal is to locate, suppress and neutralize the enemy. In our VR simulation, the two fire teams are – Team Alpha and Team Bravo. Team Bravo will be comprised of human users working together with an artificial intelligence (AI)-driven squad leader and Team Alpha, an infantry team consisting of AI-driven agents/squad members. The drill should take approximately 5 to 10 minutes to complete, and we anticipate each team will perform 3 to 5 iterations of the drill.

Data will be collected on events, movements, body gestures, head movements, log files of human-computer interactions (HCI), verbalizations, electrodermal activity, and heart rate data from the human users using a combination of GIFT and the Capturing and Logging OpenVR (CLOVR) open-source tool (Segarra Martinez et al., 2024). CLOVR is a tool for collecting data from any VR application built with the OpenVR API (Application Programming Interface), including closed-source VR consumer games and experiences. It supports capturing and logging VR device poses, VR actions, microphone audio, VR views, VR videos, and even the presentation of in-VR questionnaires. We are currently creating a new version of CLOVR that is compatible with other VR software development kits (SDKs) aside from the OpenVR SDK, such as the Meta SDK, which we call the Multimodal Observational OpenVR (MOOVR) Toolkit. This will allow us to also capture eye tracking data by using the Meta Quest Pro headset in conjunction with the MOOVR Toolkit. All of the data captured with MOOVR will be made available to the GIFT framework to afford long-term data tracking on an individual user basis. Below, we discuss more in detail about the multimodal metrics that will be further incorporated within our GIFT implementation as a multidimensional measure of teamwork metrics at both the individual and the team levels.

## **THEORETICAL FRAMEWORK**

---

Simulation based training has benefited from the ability to automatically evaluate learner performance using multimodal data instead of relying on manual analysis by domain expert review (Azevedo & Wiedbusch, 2023; Biswas et al., 2020; Goldberg et al., 2021; Vatrál et al., 2022). This is typically accomplished by capturing traces of user behaviors during the simulation to make inferences about learning and cognitive, behavioral, affective, and metacognitive processes based on theory (Winne & Azevedo 2022). In addition to traditional audio and video data of the learner going through the simulation-based training environment, other objective measures can be captured using eye-tracking, gesture-recognition, and physiological data (e.g., heart rate, electrodermal activity), in addition to subjective measures collected via self-reports and verbalizations. However, all these data must be contextualized within both the task and a theoretical model of the performance metrics.

Our work is an extension of the hierarchical Affective, Behavioral, and Cognitive model of teamwork (H-ABC; Vatrál et al., 2022). According to this model, teamwork is comprised of a series of temporally dynamic affective (e.g., mutual trust, self-efficacy), behavioral (e.g., communication, coordination), and cognitive (e.g., team mental models, team learning) processes. These processes can be organized into a multi-level hierarchical structure in which more high-level abstract teamwork processes can be directly linked to low-level directly observable skills and competencies specific to a context or domain. This model continues to be updated and improved upon through iterative empirical work to allow for more explicit mapping of measures to context-specific skills, knowledge, and abilities. However, this model does not explicitly include metacognitive processes currently. As previously described, metacognition, or the monitoring and regulation of first-order cognition, is essential for individuals to make evaluations and reflections of their performance to modify and adapt to changing conditions and standards (Katyal & Fleming, 2024). Group metacognition can strengthen the accuracy of this monitoring through discussion of experiences, perceptions, and evaluations especially in the absence of objective feedback (Wolfe, 2018).

Below, we describe our extension to the H-ABC framework to include metacognition at the highest abstraction level (among affect, behavior, and cognition), several mid-level processes (i.e., planning,

evaluation, and reflection), and several context-specific low-level skills and competencies (e.g., setting goals, identifying gaps in task understanding). This expansion pulls from multiple theories of metacognition (e.g., Greene & Azevedo, 2009; Lobczowski, 2022) and socially shared regulation of learning (e.g., Järvelä et al., 2023).

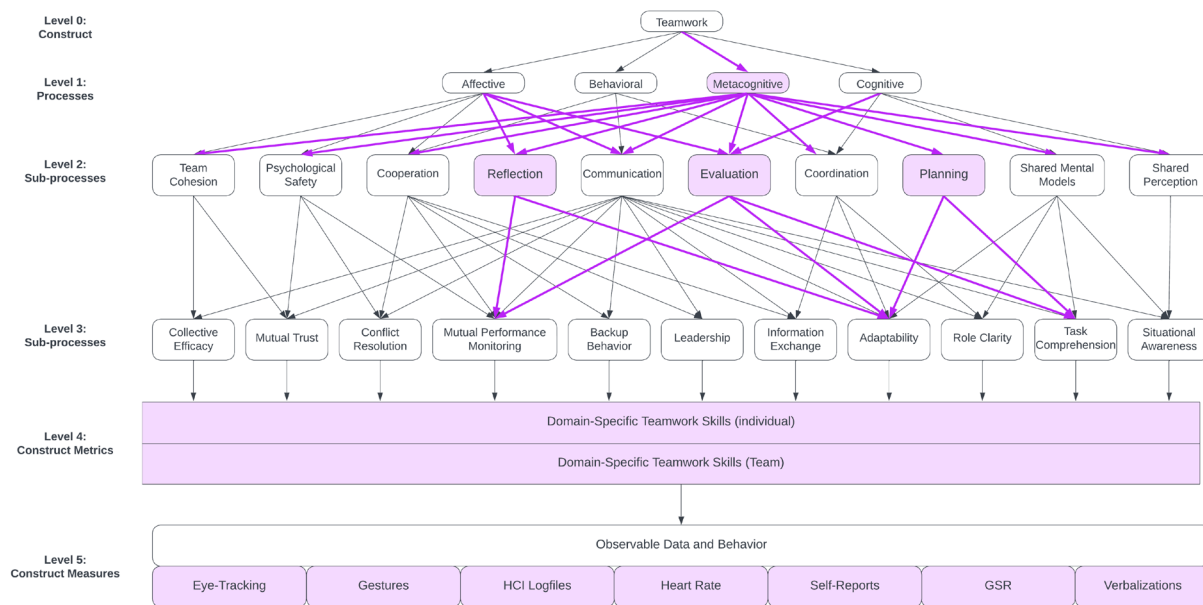
### Extending H-ABC to Integrate Metacognitive Team Collaboration Processes

Our extension focuses on the addition of metacognitive processes; however, we also slightly adapt levels 4 and 5 to our current study to highlight the distinction between individual and team metrics and the specific multimodal data measures of those metrics respectively. According to this extension, we consider metacognition as a level 1 process akin to affect, behavior, and cognition. While by definition, metacognition includes the regulation of first order cognition, we have refrained from making any interactions across level 1 processes explicit. Instead, we consider these interactions through their joint influence on subprocesses. For example, we define “Team Cohesion”, a level 2 sub-process, as a metacognitive-affective process. We show that metacognition theoretically exhaustively influences each of the defined level 2 sub-processes (see Table 1) based on various literature in the field.

**Table 1. Literature exemplars theoretically justifying each Metacognitive -> Level 2 sub-processes connection**

Level 2 Sub-Process	Description	Exemplar Literature
Team Cohesion	The shared multidimensional desire/bond that drives teams to want to work and stay together that includes common/shared tasks and goals, social relationships, sense of belongingness, group pride, and morale (Salas et al., 2015)	Garrison, 2022 Kozlowski & Chao, 2012; Lobczowski et al., 2021;
Psychological Safety	The evaluation of how “safe” individuals within a team feel in bringing up certain subjects or seeking assistance (Edmondson, 1999)	Dibble et al., 2019; Thompson & Cohen, 2012 Tucker et al., 2006
Cooperation	A structure for joint interaction towards a defined task or goal (Panitz, 1996)	Cheong, 2010; Nonose et al., 2014; Stevens et al., 2016
Communication	The explicit expression of ideas through words, actions, and facial expressions (Dillenbourg & Traum, 2006)	Carlson, 2016; Folomeeva & Klimochkina, 2021; Joksimovic et al., 2020
Coordination	The dynamics of team member interaction and the environmental dynamics they are acting within under a shared mental model (Gorman et al., 2010)	Thompson & Cohen, 2012; Keestra, 2017; Kwon et al., 2013;
Shared Mental Models	The team’s internal representation and cognitive structure of their task, team, interactions, and environment (Jonker et al., 2010)	Gorman et al., 2010; Thomspson & Cohen, 2012; Mohammed et al., 2017
Shared Perception	The symmetrical awareness of each individual’s understanding of their environment including any unique affordances or limitations due to incomplete information or abilities (Matarese et al., 2022)	Gormon et al., 2010; Jamil et al., 2023; Järvelä & Hadwin, 2013

Additionally, we have included three new level 2 sub-processes – (1) reflection, (2) evaluation, and (3) planning (Greene & Azevedo, 2009; Lobczowski, 2022). Reflection refers to the monitoring of one’s cognition about their practice (i.e., behavior or approach) to make adjustments (McAlpine et al., 1998). Reflection, in our context, therefore, is directly related to the monitoring of team performance. Evaluation refers to the monitoring and appraisal of one’s affect, behavior, and cognition relative to plans and goals (Greene & Azevedo, 2009; Lobczowski, 2022) which may then trigger future changes in response. We have theoretically tied evaluation to the mutual performance monitoring, adaptability, and task comprehension level 3 sub-processes. Evaluation is vital for adaption to changing environmental factors and team dynamics. For example, if we are unable to recognize a changing affective atmosphere in response to an environmental change (e.g., failing to recognize growing group frustration or heightened arousal during an ambush), we may then fail to act accordingly (e.g., emotionally regulate to avoid impulsive behavior) resulting in lowered team performance. Finally, planning refers to the coordination of selecting cognitive processes that once executed behaviorally will result in a change in state towards a set of (sub)goals (Greene & Azevedo, 2009). While traditionally we think of planning as happening only prior to any task performance, planning can happen intermediately throughout a task as an individual evaluates and reflects on their current state before choosing next steps. As such, we have directly tied the level 3 subprocesses of adaptability and task-comprehension to planning.



**Figure 2. The revised H-ABC model for evaluation of teamwork behaviors, as developed in Vatrul, et al. (2022) to include metacognitive processes (additions highlighted in purple).**

In addition to the new metacognitive sub-processes on levels 2 and 3, our extension highlights a distinction within the level 4 construct metrics of teamwork. Specifically, we identify that under our context we will have metrics for both the individual and the team. Teamwork by a team involves more than just a collection or aggregation of simultaneous coordinated individual actions, but rather may be considered an emergence of coordination and joint actions (Cohen & Levesque, 1991; Gorman et al., 2017). That is, it is a dynamical system (Gorman et al., 2017) that requires advanced modeling approaches using a multimodal data approach to measuring the various metrics at both the individual and team (i.e., system) level. We have identified the data sources we will be using in our context within level 5 and provide a brief explanation of these measures below.

## Multimodal Measures of Metacognition in Team Collaboration

As we have previously established, VR simulations are positioned to provide rich traces of affective, behavioral, cognitive, and metacognitive processes that are traditionally inferred in observation-only based assessment of team performance and teamwork. Multimodal trace data is highly valuable in its ability to provide unobtrusive insights into various psychological constructs and processes as they unfold in real time (Azevedo & Gasevic, 2019). Furthermore, having multiple streams (or sources) of data can allow us to combine and fuse across modalities to provide more context-rich data and interpretations than a singular channel can provide alone (Wiedbusch et al., 2023). This holds especially true for the black box that is cognition and metacognition in which these processes must be inferred from observable behaviors (Azevedo & Wiedbusch, 2023).

Adapted from the learning analytics field (Ochoa, 2022), we will follow a similar construct mapping process in which each of our level 4 psychological constructs (e.g., “squad member identifies gap in task understanding”) are mapped to observable behaviors (e.g., verbalizations between team members, gross level movement away from objectives, head tilts or prolonged examination of objective instructions). These behaviors collected via a suite of available multimodal data captured and synchronized using the MOOVR Toolkit including environmental or user events, movements, body gestures, head movements, log files of HCI, verbalizations, electrodermal activity, and heart rate data. Within each of these behaviors are multiple analytics or metrics that can be compiled. For example, it could be the frequency of task questioning utterances (e.g., “I don’t know what I am supposed to do here? What are we doing? What needs to be done?”) or the dwell time on instructions or team leader providing instructions (and the associated deviation from the average expected dwell time). It is important to note that there can be multiple behaviors associated with each teamwork construct and multiple analytics that can be derived from each observable behavior. After our first study, our team will be examining the optimal number of these behaviors and analytics that are required to best capture and model each construct to help reduce the number of dimensions and analytical resources required from this approach.

## IMPLEMENTATION IN GIFT

---

Next, we briefly outline our proposed implementation of our integrated feedback framework in GIFT. The work described in this paper is planned to be integrated into GIFT after it has been developed. There are two approaches in which it can be applied in GIFT: technical and theoretical. From a technical perspective, GIFT has previously been integrated with Unity, which will allow for information to be passed between the Battledrill 2A scenario that is in development and the gateway module in GIFT. The assessment of performance during the scenario can be implemented through the scenario itself as well as in a domain knowledge file (DKF) in GIFT. It is anticipated that some of the generalizable assessments, and relevant condition classes which assess behaviors could become part of the standard condition classes included with GIFT. From a theoretical perspective, the development of GIFT has continually been rooted in theory. The initial H-ABC model (Vatral et al., 2022) continues to provide a theoretical basis that can be applied in GIFT, and the addition of metacognition adds a new dimension that can also be tracked and utilized for adaptivity in GIFT.

## CONCLUSIONS AND FUTURE DIRECTIONS

---

Effective implementation of automatic feedback in an intelligent tutoring system demands a strong theoretical grounding to make constructive interpretations of multimodal data signals for inherently noisy and complex tasks that involve large teams. In this paper, we describe a theoretical expansion of a model of teamwork to include metacognition (at the individual and group level) while exploring what type of data

may be best for the collection of these processes' traces during simulation-based training. Metacognitive processes of planning, evaluating, and reflecting are what drive team adaptability to complex and rapidly shifting environments. Their inclusion within models of teamwork is inherently messy due to the entanglement of metacognition, cognition, affect, and behavior. However, we argue that the inclusion and acknowledgement of these processes for application far outweigh the loss of some distinction between process origination at higher levels within the model.

In addition, we discussed the future implementation of our multimodal measures of teamwork metrics within GIFT to be used as an assessment and feedback tool for simulation-based training. By offering data outside of traditional subjective expert graded performance feedback, we anticipate learners will be able to garner a more holistic view of not only their individual but team level performance to target specific skills, behaviors, and strategies in future iterations of training. Future work on this project will include the development of an analytical after action review to be integrated with GIFT to provide actionable feedback and performance review.

While this work above is in the initial theoretical development of expanding a framework, we are also designing a series of empirical studies by which to refine this framework and inform future work. These studies will be conducted to train our models of team dynamics and performance and assess the impact of various multimodal signals as indicators of teamwork during simulation-based training. Based on the outcome of these future models, we will continue the development of a VR environment with adaptive and intelligent artificial agents capable of performing as a team with human counterparts to help reduce cost and supplement current approaches to team training.

## ACKNOWLEDGEMENTS

---

Research was sponsored by DEVCOM-SC-SED-TSD and was accomplished under Cooperative Agreement Number W912CG-23-2-0004. 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 DEVCOM-SC-SED-TSD or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein.

## REFERENCES

---

- Azevedo, R. (2020). Reflections on the field of metacognition: Issues, challenges, and opportunities. *Metacognition and Learning* 15, 91-98.
- Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior*, 96, 207-210.
- Azevedo, R., & Wiedbusch, M. (2023). Theories of metacognition and pedagogy applied to AIED systems. In B. du Boulay, A. Mitrovic & K. Yacef (Eds.), *Handbook of artificial intelligence in education* (pp. 45-67). Edward Elgar Publishing.
- Azevedo, R., Taub, M., Mudrick, N. V., Martin, S. A., & Grafsgaard, J. (2018). Using multi-channel trace data to infer and foster self-regulated learning between humans and advanced learning technologies. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 254-270). Routledge.
- Balint, B. N., Stevens, B., Dudfield, H., & Powell, W. (2020, November). Exploring the characteristics of immersive technologies for teamwork. Paper presented at *Virtual Interservice/Industry Training, Simulation, and Education Conference*, Orlando, Fl. 2020.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Biswas, G., Rajendran, R., Mohammed, N., Goldberg, B. S., Sottolare, R. A., Brawner, K., & Hoffman, M. (2020). Multilevel learner modeling in training environments for complex decision making. *IEEE Transactions on Learning Technologies*, 13(1), 172–185.
- Carlson, E. (2016). Meta-accuracy and relationship quality: Weighing the costs and benefits of knowing what people really think about you. *Journal of Personality & Social Psychology*, 111(2), 250-264.
- Cheong, C. (2010). From group-based learning to cooperative learning: A metacognitive approach to project-based group supervision. *Informing Science: The International Journal of an Emerging Transdiscipline*, 13, 73-86.
- Cohen, P. R., & Levesque, H. J. (1991). Teamwork. *Nous*, 25(4), 487-512.
- Cooke, N. J., Gorman, J. C., Winner, J. L., & Durso, F. T. (2007). Team cognition. In F. T. Durso, R.S. Nickerson, S. T. Dumais, S. Lewandowsky, & T. J. Perfect (Eds.), *Handbook of applied cognition*, (2<sup>nd</sup> ed., pp. 239-268). Wiley.
- Dibble, Linda S. Henderson & Zachary C. Burns (2019) The impact of students' cultural intelligence on their psychological safety in global virtual project teams, *Journal of Teaching in International Business*, 30(1), 33-56.
- Dillenbourg, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *Journal of the Learning Sciences*, 15(1), 121-151.
- Dorn, A. W., Webb, S., & Pâquet, S. (2020). From wargaming to peacegaming: Digital simulations with peacekeeper roles needed. *International Peacekeeping*, 27(2), 289-310.
- Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2), 350–383.
- Endsley, M. R., & Jones, D. G. (2001). Disruptions, interruptions and information attack: impact on situation awareness and decision making. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 45, Issue 2, pp. 63-67). SAGE Publications.
- Flavell, J. H. (1976). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10), 906.
- Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75, 241-268.
- Folomeeva, T. V., & Klimochkina, E. N. (2021). Social metacognition in the process of decision making. In C. Pracana & M. Wang Eds.) *Psychological applications and trends 2021*, (pp. 277-281). inScience Press.
- Garrison, D. R. (2022). Shared metacognition in a community of inquiry. *Online learning*, 26(1), 6-18.
- Goldberg, B., Owens, K., Hellman, K., Robson, R., Blake-Plock, S., Hoffman, M., & Gupton, K. (2021). Forging competency and proficiency through the synthetic training environment with an experiential learning for readiness strategy. In *Interservice/Industry Training, Simulation, and Education Conference (IITSEC)*, Orlando, FL. 2021.
- Gorman, J. C., Amazeen, P. G., & Cooke, N. J. (2010). Team coordination dynamics. *Nonlinear Dynamics, Psychology, and Life Sciences*, 14(3), 265.
- Gorman, J. C., Dunbar, T. A., Grimm, D., & Gipson, C. L. (2017). Understanding and modeling teams as dynamical systems. *Frontiers in psychology*, 8, doi: 10.3389/fpsyg.2017.01053
- Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary Educational Psychology*, 34(1), 18-29.
- Jamil, A. F., Siswono, T. Y. E., & Setianingsih, R. The emergence and form of metacognitive regulation: Case study of more and less successful outcome groups in solving geometry problems collaboratively. *Mathematics Teaching Research Journal*, 15(1), 25-43.



## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Järvelä, S., & Hadwin, A. F. (2013). New frontiers: Regulating learning in CSCL. *Educational Psychologist, 48*(1), 25-39.
- Järvelä, S., Nguyen, A., Vuorenmaa, E., Malmberg, J., & Järvenoja, H. (2023). Predicting regulatory activities for socially shared regulation to optimize collaborative learning. *Computers in Human Behavior, 144*, 107737.
- Johnston, J. H., Phillips, H. L., Milham, L. M., Riddle, D. L., Townsend, L. N., DeCostanza, A. H., Patton, D. J., Cox, K. R., & Fitzhugh, S. M. (2019). A team training field research study: Extending a theory of team development. *Frontiers in Psychology, 10*, 1480. <https://doi.org/10.3389/fpsyg.2019.01480>
- Johnston, J. H., Sottolare, R. A., Kalaf, M., & Goodwin, G. (2022). Training for team effectiveness under stress. In A. Sinatra, A. C. Graesser, X. Hu, B. Goldberg, A. J. Hampton, & J. H. Johnston (Eds.), *Design Recommendations for Intelligent Tutoring Systems: Volume 9-Competency-Based Scenario Design* (pp. 69-74). Army Research Laboratory.
- Joksimovic, S., Marshall, R., Kovanovic, V., Ladjal, D., James, N., & Pardo, A. (2020). The importance of metacognitive regulation for work-integrated learning. *Australian Collaborative Education Network (ACEN) 2020 Conference*. Melbourne, Australia (online).
- Jonker, C. M., Van Riemsdijk, M. B., & Vermeulen, B. (2010, August). Shared mental models: A conceptual analysis. In M. De Vos, N. Fornara, J. V. Pitt, & G. Vouros (Eds.), *International workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems* (pp. 132-151). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Katyal, S., & Fleming, S. M. (2024). The future of metacognition research: Balancing construct breadth with measurement rigor. *Cortex, 171*, 223-234.
- Keestra, M. (2017). Metacognition and reflection by interdisciplinary experts: Insights from cognitive science and philosophy. *Issues in Interdisciplinary Studies, 35*, 121 – 169.
- Kozlowski, S. W. J. (2018). Enhancing the effectiveness of work groups and teams: A reflection. *Perspectives on Psychological Science, 13*(2), 205–212.
- Kozlowski, S. W., & Chao, G. T. (2012). The dynamics of emergence: Cognition and cohesion in work teams. *Managerial and Decision Economics, 33*(5-6), 335-354.
- Kwon, K., Hong, R. Y., & Laffey, J. M. (2013). The educational impact of metacognitive group coordination in computer-supported collaborative learning. *Computers in Human Behavior, 29*(4), 1271-1281.
- Lobczowski, N. G. (2022). Capturing the formation and regulation of emotions in collaborative learning: The FRECL coding procedure. *Frontiers in Psychology, 13*, 846811.
- Lobczowski, N. G., Lyons, K., Greene, J. A., & McLaughlin, J. E. (2021). Socially shared metacognition in a project-based learning environment: A comparative case study. *Learning, Culture and Social Interaction, 30*, 100543.
- Martín-Hernández, P., Gil-Lacruz, M., Gil-Lacruz, A. I., Azkue-Beteta, J. L., Lira, E. M., & Cantarero, L. (2021). Fostering university students' engagement in teamwork and innovation behaviors through game-based learning (GBL). *Sustainability, 13*(24), 13573.
- Matarese, M., Rea, F., & Sciutti, A. (2022). Perception is only real when shared: A mathematical model for collaborative shared perception in human-robot interaction. *Frontiers in Robotics and AI, 9*, 733954.
- McAlpine, L., Weston, C., Beauchamp, C., Wiseman, C., & Beauchamp, J. (1998). Building a metacognitive model of reflection. *Higher education, 37*, 105-131.
- Meslec, N., Duel, J., Soeters, J. (2020). The role of teamwork on team performance in extreme military environments: An empirical study. *Team Performance Management, 26*(5), 325-339.
- Mohammed, S., Hamilton, K., Sánchez-Manzanares, M., & Rico, R. (2017). Team cognition: Team mental models and situation awareness. In E. Salas, R. Rico, & J. Passmore (Eds.) *The Wiley Blackwell handbook of the psychology of team working and collaborative processes*, 369-392. Wiley Blackwell, Hoboken, NJ.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Nelson, T., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125-173.
- Nonose, K., Kanno, T., & Furuta, K. (2014). Effects of metacognition in cooperation on team behaviors. *Cognition, Technology & Work*, 16, 349-358.
- Norman, E., Pfuhl, G., Sæle, R. G., Svartdal, F., Låg, T., & Dahl, T. I. (2019). Metacognition in psychology. *Review of General Psychology*, 23(4), 403-424.
- Ochoa, X., (2022). Multimodal learning analytics-Rationale, process, examples, and direction. In X. Ochoa, C. Lang, G. Siemens, A. Wise, D. Gasevic, & A. Merceron (Eds.) *The handbook of learning analytics*, (2<sup>nd</sup> Ed., pp. 54-65). SoLAR, Vancouver, BC.
- Panitz, T. (1996) A definition of Collaborative vs. Cooperative Learning. *Deliberations*, 1-3.
- Salas, E., Grossman, R., Hughes, A. M., & Coultas, C. W. (2015). Measuring team cohesion: Observations from the science. *Human Factors*, 57(3), 365-374.
- Segarra Martinez, E., Malik A. A. & McMahan, R. P. CLOVR: Collecting and logging OpenVR data from SteamVR applications (in Press). In *Proceedings of 2024 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*, Orlando, FL, USA, 2024, pp. 485-492, doi: 10.1109/VRW62533.2024.00095.
- Sottolare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). The generalized intelligent framework for tutoring (GIFT). Orlando, FL: US Army Research Laboratory–Human Research & Engineering Directorate (ARL-HRED).
- Sottolare, 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, 1-19.
- Stevens, C. A., Taatgen, N. A., & Cnossen, F. (2016). Instance-based models of metacognition in the prisoner's dilemma. *Topics in Cognitive Science*, 8(1), 322-334.
- Tarricone, P. (2011). *The Taxonomy of Metacognition*. Hove: Psychology Press.
- Thompson, L., & Cohen, T. R. (2012). Metacognition in teams and organizations. In P. Briñol, & K. DeMarree (Eds.), *Social metacognition* (pp. 283-302). Psychology Press.
- Tucker, A. L., Nembhard, I. M., & Edmondson, A. (2006). Implementing new practices: An empirical study of organizational learning in hospital intensive care units. *Management Science*, 53, 894-907.
- Vatral, C., Biswas, G., & Goldberg, B. S. (2022). Multimodal learning analytics using hierarchical models for analyzing team performance. In A. Weinberger, W. Chen, D. Hernández-Leo, & B. Chen (Eds.), *Proceedings of the 15th International Conference on Computer Supported Collaborative Learning-CSCL 2022* (pp. 403-406). International Society of the Learning Sciences.
- Wiedbusch, M., Dever, D., Li, S., Amon, M. J., Lajoie, S., & Azevedo, R. (2023). Measuring multidimensional facets of SRL engagement with multimodal data. In V. Kovanovic, R. Azevedo, D. C. Gibson, & D. Ifenthaler (Eds.), *Unobtrusive observations of learning in digital environments: Examining behavior, cognition, emotion, metacognition and social processes using learning analytics* (pp. 141-173). Cham: Springer International Publishing.
- Winne, P. H. (2018). Learning analytics for self-regulated learning. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2<sup>nd</sup> ed, pp. 36-48). Routledge.
- Winne, P., & Azevedo, R. (2022). Metacognition and self-regulated learning. In R. K. Sawyer (ed.), *The Cambridge handbook of the learning sciences* (pp. 93-113). Cambridge University Press.
- Wolfe, W. (2018). Poor metacognitive awareness of belief change. *Quarterly Journal of Experimental Psychology*, 71(9), 1898-1910

## ABOUT THE AUTHORS

---

**Megan Wiedbusch, Ph.D.** is a postdoctoral researcher at the School of Modeling, Simulation, and Training at the University of Central Florida. Her research is focused on the measurement of the dynamics of metacognition and engagement using traditional (i.e., self-reports) and unobtrusive multimodal (e.g., eye tracking, facial expressions, log files) methodological and analytical approaches across contexts (e.g., health care, K-12 education, teacher training) and learning environments (e.g., VR, simulations, ITS, and GBLEs). She conducts laboratory, classroom, and in-situ studies to model human (meta)cognition and behavior during complex learning to inform the design of human-centered intelligent learning and training technologies.

**Ryan P. McMahan, Ph.D.** is an Associate Professor of Computer Science at the University of Central Florida (UCF). He directs the eXtended Reality & Training (XRT) Lab, which focuses on using extended reality (XR) and virtual reality (VR) technologies to facilitate and enhance training and education. Dr. McMahan is a National Science Foundation (NSF) CAREER Award winner, and his research has been funded by multiple NSF grants, Defense Advanced Research Projects Agency (DARPA) projects, and the U.S. Army Research Laboratory. At UCF, he is the Associate Program Director for the Mixed Reality Engineering Graduate Certificate program and won the sole Excellence in Graduate Teaching Award for the College of Engineering and Computer Science in 2023. Dr. McMahan received his Ph.D. in Computer Science and Applications from Virginia Tech in 2011.

**Anne M. Sinatra, Ph.D.** is a Research Psychologist at U.S. Army Combat Capabilities Development Command Soldier Center, Simulation & Training and Technology Center in Orlando, FL. She has a Ph.D. in Applied Experimental and Human Factors Psychology from the University of Central Florida (UCF). Her research focuses on applying cognitive psychology and human factors principles to computer-based education and adaptive training to enhance learning. She is a member of the research team for the award winning Generalized Intelligent Framework for Tutoring (GIFT).

**Benjamin Goldberg, Ph.D.** is a senior research scientist at the U.S. Army Combat Capability Development Command–Soldier Center, and is co-creator of the Generalized Intelligent Framework for Tutoring (GIFT). Dr. Goldberg is the technical lead for a research program focused on the development and evaluation of Training Management Tools for future Army training systems. His research is focused on the application of intelligent tutoring and artificial intelligence techniques to build adaptive training programs that improve performance and accelerate mastery and readiness. Dr. Goldberg has researched adaptive instructional systems for the last 15 years and has been published across several high-impact proceedings. He holds a Ph.D. in Modeling & Simulation from the University of Central Florida.

**Lisa N. Townsend** is a Senior Research Psychologist at the U.S. Army Combat Capabilities Development Command Soldier Center, Simulation & Training Technology Center. She has an M.S. in Industrial/Organizational Psychology and a B.A. in Psychology, from the University of Central Florida (UCF). She has worked on many diverse teams including those within Research and Development, Technology Transfer, Instructional Systems Design, and Human Systems Integration. Ms. Townsend's areas of expertise involve team training, Front End Analysis (FEAs), Training Systems Analyses (TSAs), Instructional Systems Design (ISD), Training Effectiveness Evaluations (TEEs), and the development of training and organization related metrics. Her efforts in these areas have spanned across Services and platforms.

**Joseph J. LaViola Jr., Ph.D.** is the Charles N. Millican Professor of Computer Science and directs the Interactive Computing Experiences Research Cluster at the University of Central Florida. He is also a visiting scholar in the Computer Science Department at Brown University. He is the former director of the Modeling and Simulation graduate program at UCF. His primary research interests include pen- and touch-based interactive computing, virtual and augmented reality, 3D spatial interfaces, human-robot interaction, multimodal interaction, and user interface evaluation. He has published over 185 refereed journal and conference papers, 8 book chapters, and has 5 patents. His work has appeared in journals such as ACM TIIS, ACM TOCHI, IEEE PAMI, Presence, and IEEE Computer Graphics & Applications, and he has presented research at conferences including ACM CHI, ACM IUI, IEEE Virtual Reality, and ACM SIGGRAPH. He is also the lead author on the second edition of "3D User Interfaces: Theory and Practice", the first comprehensive book on 3D user interfaces. In 2009, he won an NSF Career Award to conduct research on mathematical sketching. Joseph received a Sc.M. in Computer Science in 2000, a Sc.M. in Applied

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

*Mathematics in 2001, and a Ph.D. in Computer Science in 2005 from Brown University. He is a senior member of the ACM and IEEE.*

**Roger Azevedo, Ph.D.** *is a professor at the School of Modeling Simulation and Training at the University of Central Florida. He is also an affiliated faculty in the Departments of Computer Science and Internal Medicine at the University of Central Florida and the lead scientist for the Learning Sciences Faculty Cluster Initiative. His main research area includes examining the role of cognitive, metacognitive, affective, and motivational self-regulatory processes during learning with advanced learning technologies (e.g., intelligent tutoring systems, hypermedia, multimedia, simulations, serious games, immersive virtual learning environments, human digital twins). He has published over 300 peer-reviewed papers and chapters and has refereed conference proceedings in educational, learning, cognitive, and computational sciences. He is the co-editor of the British Journal of Educational Psychology and serves on the editorial boards of several top-tiered interdisciplinary journals. He is a fellow of the American Psychological Association and the recipient of the prestigious Early Faculty Career Award from the National Science Foundation.*



**THEME III:  
ARTIFICIAL INTELLIGENCE  
APPLICATIONS**



# Leveraging TCAT for Advanced Team Communication Analysis and Performance Assessment in GIFT

Randall Spain<sup>1</sup>, Wookhee Min<sup>2</sup>, Nicholas Roberts<sup>3</sup>, Vikram Kumaran<sup>2</sup>, Jay Pande<sup>2</sup>, and James Lester<sup>2</sup>  
U.S. Army Combat Capabilities Development Command, Soldier Center, Training and Simulation Division<sup>1</sup>, North Carolina State University<sup>2</sup>, Dignitas Technologies<sup>3</sup>

## INTRODUCTION

---

Recent advances in natural language processing (NLP) are offering new opportunities to support team communication analysis (Jensen et al., 2021; Min et al., 2021; Pande, Min, et al., 2023). When integrated into adaptive instructional systems (AISs), these capabilities can allow instructors, leaders, and units to obtain timely insights into teamwork and team performance through multi-faceted analysis of team dialogue. This paper discusses ongoing efforts associated with the Team Communication Analysis Toolkit (TCAT) project to integrate team communication analytics into the Generalized Intelligent Framework for Tutoring (GIFT). TCAT is an NLP-driven framework that analyzes team communication data using machine learning models to assign descriptive dialogue labels to each utterance and extract salient patterns in the team’s communication (Spain et al., 2022). In particular, TCAT utilizes deep learning-based NLP models for encoding natural language and inferring latent information such as dialogue acts and information flow labels from team communication data. TCAT also offers a user interface for investigating team communication data; the interface includes a set of data visualization techniques that effectively demonstrate salient communication patterns associated with team members’ roles and utterances.

We begin by discussing the latest enhancements to TCAT’s NLP architecture (Pande, Min et al., 2023). Building on the team’s previous efforts, the project team has accomplished several major tasks this year to advance TCAT’s capabilities and support near-real-time assessment of team dialogue during training events. In particular, we discuss integrating T5 (text-to-text transfer transformer) models (Raffel et al., 2020) into the TCAT software. This integration yields significant improvements in TCAT’s dialogue classification capabilities, outperforming the previous state-of-the-art models in classification accuracy for both dialogue act recognition and information flow classification tasks on the Squad Overmatch dataset, streamlining multi-task learning and demonstrating enhanced generalization. In addition, we report updates on our goal to incorporate real-time automatic speech recognition (ASR) features in TCAT. This implementation utilizes a multi-thread approach in the software to support users to dynamically start and stop their recordings as the training event takes place. This module was incorporated into TCAT’s latest software, and the feature is available in the software. Further, we discuss how we have expanded TCAT to support crew gunnery communication analysis within fire teams during target engagement exercises. Using a set of human-labeled transcripts, we designed and developed two approaches, including  $n$ -gram-based classification and Sentence-BERT (SBERT) embedding-based classification. We share preliminary findings on fire command classification results based on this initial investigation.

Finally, we discuss progress towards integrating TCAT within GIFT. This includes an overview of several major development milestones, including (1) the creation of a Gateway module interface, *TCATInterface*, making TCAT a registered training application usable within GIFT’s course creator, (2) the establishment of a data pipeline that allows for GIFT to communicate with TCAT over a TCP socket connection and (3) the creation of a new *SpeakPhrase* condition class for transcript assessments in GIFT that allows the author to define a sequence of words (or a regular expression) that TCAT’s ASR-generated transcript is expected to contain and provides an “At Expectation” assessment if the transcript matches. We conclude by discussing plans to extend our assessment approach to include additional dialogue analysis outputs such as dialogue acts and information flow labels from TCAT to support content-based team coordination measurement.

## BACKGROUND

---

Analyzing verbal communication among team members offers valuable insights into the behaviors and interactions that impact team performance (Marlow et al., 2018). Advances in NLP have enabled the automation of transcript generation and subsequent syntactic and semantic analyses of dialogue. TCAT (Min et al., 2021; Pande, Min et al., 2023) leverages these advancements to facilitate the analysis of spoken communication between team members. The NLP pipeline for TCAT consists of two primary components: ASR and automated utterance labeling. TCAT uses Microsoft Azure’s Speech-to-Text cloud service for ASR, converting spoken communication into text for downstream NLP tasks. For utterance labeling, TCAT uses a pre-trained model to perform multitask classification, labeling each utterance with one of nine *dialogue acts*, which reflect the intent of the speech (e.g., PROVIDE INFORMATION, ACKNOWLEDGEMENT, COMMAND), and, if applicable, one of 18 *information flow* labels, which reflects how information was passed between different levels of the team’s hierarchy (e.g., PROVIDE INFORMATION DOWN when the squad or team leader shares information with team members, REQUEST INFORMATION UP when team members ask for information from the squad or team leader). The dialogue act and information flow recognition model within this TCAT was trained using labeled transcripts from the Squad Overmatch Mission 3 dataset, comprising 6,181 utterances captured from six squads during a 45-minute live training exercise (Johnston et al., 2019; Pande, Min et al., 2023). TCAT offers users valuable statistics and data visualization techniques to support the analysis of team communication.

Since our initial work on TCAT, advancements in deep neural architectures for NLP have led to significant performance improvements across various tasks. The Transformer architecture, introduced by Vaswani et al. (2017), has achieved significant improvements in NLP tasks compared to previous deep learning approaches. Its key innovation is the self-attention mechanism that allows the model to focus on relevant parts of the input sequence. This mechanism is crucial for effectively modeling long sequences and preserving long-range dependencies, unlike recurrent neural networks (RNNs) and their variants (e.g., LSTMs, GRUs), which process tokens sequentially and often lose early-captured characteristics by the time they reach the end of the sequence. Transformer-based models that effectively capture long-range dependencies in long sequences have demonstrated state-of-the-art results on various benchmark datasets for tasks like machine translation, question answering, and language modeling. The key advantages of using Transformer-based models like T5 include end-to-end training and inference with free-form text input and output, the ability to perform multiple tasks simultaneously, the potential for generalization to new datasets, and the use of task-specific prefixes to infer characteristics of tasks from the phrasing of prompts. Along with significant advancements in NLP capabilities, over the past year the project team has expanded TCAT’s features to support more accurate dialogue act recognition and improved information flow classification to facilitate deeper analysis of team communication patterns within and between teams in a squad. Next, we discuss the improvements to TCAT’s architecture and our work towards expanding TCAT’s NLP models to support team communication analysis in additional domains.

## DEVELOPMENT UPDATES

---

### Integration of T5 Models into TCAT software

Last year, we discussed the benefits of the T5 architecture, how we created separate T5 models (i.e., two single-task models) for dialogue act recognition and information flow classification, and how we tested the classification performance using transcripts from the Squad Overmatch project (Pande, Paul et al., 2023). This year, the TCAT project team integrated a multitask T5 model into the TCAT software that performs both dialogue act and information flow labeling simultaneously. The integration of T5 not only streamlined multi-task learning but also demonstrated remarkable generalization capabilities across different



distributions of dialogues. To facilitate broader usage, the team also developed a Windows build of the TCAT software using PyInstaller, enhancing the accessibility and usability of the application.

The results of cross-validation based on data from five out of six squads show that our T5 framework significantly surpasses the baselines using the majority class (accuracies of 25.7% for dialogue act recognition and 45.8% for information flow classification) for the Squad Overmatch dataset (Pande, Min et al., 2023). The most effective model for dialogue act recognition utilized a learning rate of  $3e-4$ , a batch size of 8, and warm-up steps amounting to 10% of the training data, achieving accuracy rates between 70.97% and 77.41% across all folds (mean = 73.55%). Similarly, the optimal model for information flow classification, which used a learning rate of  $1e-4$ , a batch size of 4, and an identical number of warm-up steps, recorded accuracies ranging from 58.62% to 71.30% across all folds (mean = 65.72%).

Using the identified optimal hyperparameter settings, new models were trained on the entire training dataset and tested on a held-out test set (data from the remaining one squad out of six squads), which remained unseen during the cross-validation to prevent data leakage. For the inner split of the training data, 90% was used for training and 10% was used for validation to guide the early stopping process. Accuracy on this test set was higher than cross-validation averages at 75.92% for dialogue act recognition and 67.23% for information flow classification, outperforming previous best performances using this dataset based on conditional random field models utilizing ELMo embeddings (Min et al., 2021). Furthermore, these improvement margins over the majority-class baselines were substantial: 195.4% for dialogue act recognition and 46.8% for information flow classification.

Our framework was also tested on a domain-transfer task, where T5 models were trained on all of the six squads mentioned above and tested on dialogue act recognition using a dataset from a different team training exercise. This dataset contains a similar sequence of events but features different data distributions than the dataset utilized to train the T5 model. This model, trained with the optimal hyperparameters from the earlier task, achieved a 70.60% accuracy, demonstrating our framework's capability to generalize to a related, but different training exercise. These outcomes together suggested that our framework, powered by T5's transformer-based architecture and exposure to various NLP tasks, leads to significant performance improvements in team communication analysis tasks.

Furthermore, TCAT's NLP model has been enhanced to improve generalization using the speaker role input feature. Prior versions of TCAT linked speaker role input labels with squad positions (i.e., squad leader, alpha team leader, bravo team leader, team member, etc.). TCAT now incorporates a T5 model that employs a three-way split (high, middle, and low) as the speaker role label, rather than explicitly listing squad positions. This modification aims to enable the classification model to adapt to various multiparty dialogue domains and tasks that may not strictly align with the composition of a U.S. Army squad. The design of these updated input features for the NLP models will undergo further exploration and evaluation using other datasets, including Crew Gunnery and Squad Battle Drills, in the future.

### **Support for Near-Real-Time ASR**

In addition to updating TCAT's NLP architecture, another task was adding support for near real-time ASR in TCAT. Prior to the recent updates, TCAT performed ASR using Microsoft Azure's speech-to-text service based on previously captured team communication recordings and audio files. One limitation with this approach is that TCAT could not perform dialogue analysis in near real-time. Another limitation was that prior work required a manual annotation of speaker information for a combined audio from multiple participants, which was then analyzed by TCAT. To address this manual labeling issue, the project team explored speaker diarization—automatic detection of different speakers within an audio file. However, the performance was significantly hindered by ambient noise and overlapping voices. To better support near

real-time communication analysis we updated TCAT so that it can monitor live communication feeds from individual trainees and perform ASR in batches based on the user’s control of starting and stopping the recording. This implementation uses a multi-thread approach in the software to support users to dynamically start and stop their recordings as the training event takes place. Because users can start and stop audio recordings, an added benefit is that GIFT can be used to identify and provide speaker names and roles associated with each audio file, thereby eliminating the need for speaker diarization or manual labeling. TCAT currently processes audio stream data based on the user input for starting and stopping of the recording but can be extended to automatically capture recordings in preset time windows (e.g., every 10 seconds), that can be adjusted as needed. This improved ASR module has been incorporated into TCAT’s latest software, and the feature is available for use by selecting the Create a New Labeled Transcript from a Live Session task selection option. See Figure 1 for an example of automatic speech recognition of team communication implemented in TCAT.

```
Run ASR
If you want, create the "phrase.txt" file in the main TCAT directory, and it will be used for ASR.
Each phrase should be entered in a new line.
SESSION STARTED: SessionEventArgs(session_id=ab3c81b447a7498ab2ac778f66129bb5)
RECOGNIZED: Yeah. Hey, you guys are going to push probably to the right side. I'm going to have him push the left side.
RECOGNIZED: A 2A Just go ahead and start moving up towards building B5. That's Father Romanoff right there.
RECOGNIZED: Hey, just get you guys in some security positions.
RECOGNIZED: How you doing, Father?
RECOGNIZED: Where do you want you want to talk over here?
RECOGNIZED: How are you doing? Good. Good. Good. Good.
RECOGNIZED: The SO heard you had some information on Aleve Denarii. I'm just seeing if I can link up with him and talk to him. We
e have located our lock. I have got him in central area of the church.
RECOGNIZED: The building just after the market. Yes, yes. OK so but the Bishop. The Bishop is in town. He travels around Gorgas.
He is here. He would not mind to to safety words to.
RECOGNIZED: OK, the Bishop or or the Bishop would like to say a few words. His main office is in Doctor.
RECOGNIZED: 62 I just conducted the interview with Ablac Break.
```

Figure 1. Automatic speech recognition of team communication implemented in TCAT

### Crew Gunnery Communication Analysis in TCAT

The team expanded TCAT’s application scope by exploring fire command classification within crew gunnery communication. The team developed two classification approaches: an  $n$ -gram-based method and a deep representation method leveraging SBERT embeddings (Reimers & Gurevych, 2019). The process for classifying fire commands using a Crew Gunnery dataset involved extracting language representations using either method for each instance in the training and test sets, then determining the semantically closest training instances to infer the label of each test instance.

$N$ -grams are one of the simplest language models. An  $n$ -gram is a sequence of  $n$  words. For example, a 2-gram (bigram) is a two-word sequence like “enemy truck” or “target ahead,” and a 3-gram (trigram) is a three-word sequence like “enemy truck head” or “fire and adjust.” We used an  $n$ -gram-based method to generate utterance-level representations using a “bag-of-words” approach. To minimize the risk of misclassifying utterances with unrelated labels, we applied a threshold to each  $n$ -gram feature as a heuristic function. For trigrams, due to the specificity required for an exact match between training and test data instances, we utilized all trigram features extracted from the training dataset. For bigrams, where matches are more common, we set a threshold of 2; this means only bigram features appearing at least twice with the same label were retained in the feature set, while those appearing only once were removed. A similar approach was taken for unigrams, the most common type, where a threshold of 3 was applied. Once the  $n$ -gram features are finalized, the model first looks for trigram matches. If the incoming utterance includes one or more trigram matches to an utterance in the training set, the corresponding dialogue label(s) from the training utterance are assigned to the incoming utterance, and the model moves on to classify the next utterance. If a trigram match does not yield any results, the model looks for bigram matches, and subsequently for unigram matches. If there are multiple labels associated with an utterance for each  $n$ -gram, all labels are assigned to the utterance.

SBERT is a Transformer language model that extracts sentence-level embeddings for text, which can be used to assess the semantic similarity across sentences (Reimers & Gurevych, 2019). The fundamental concept behind SBERT is to train twin and triplet BERT network architectures on pairs or triplets of sentences. The primary goal of using SBERT in this research was to generate utterance-level embeddings and investigate semantic similarity across fire command statements in a vector space using a similarity metric (i.e., cosine similarity). An advantage of using SBERT for this task is that it can be used to assess the contextual meaning of entire sentences rather than just individual words. SBERT has also proven valuable in various natural language processing tasks, including sentence classification, semantic textual similarity, and question answering. It achieved state-of-the-art results on several benchmark datasets, demonstrating its effectiveness in capturing sentence-level meaning (Reimers & Gurevych, 2019). We used SBERT to extract distributed representations of crew communication utterances and classify them into fire command labels.

Table 1 presents a comparison of the results of the two modeling approaches in terms of fire command label outputs. Our initial comparisons indicated that the  $n$ -gram approach produced accurate outputs. However, the main drawback of this modeling approach is the ‘out-of-vocabulary’ problem, where if an utterance is not captured in the training set, it does not provide a match during the labeling process. Table 1 includes several examples of this limitation. The SBERT approach addresses this limitation by leveraging the semantic understanding capabilities inherent in the design of the model. SBERT excels at capturing semantic meaning from text by generating contextualized representations of utterances, which help the model robustly classify each utterance into a relevant fire command, even when the exact phrase does not appear in the training set, compared to naive  $n$ -gram models. This semantic understanding is crucial for tasks like dialogue act recognition where the intention behind the utterance matters. SBERT’s generalizability to unseen data is enabled by its learning of general patterns in language from a large corpus of text data during pre-training.

**Table 1. Comparison of SBERT and  $n$ -gram fire command labeling models.**

Speaker	Utterance	SBERT Fire Command Label	$N$ -gram Fire Command Label
Gunner	hey one enemy truck	Description	Description (Trigram)
Gunner	1 o’clock	Direction	Direction (Bigram)
Gunner	500 meters	Range	Range (Bigram)
Commander	fire and adjust	Execution Command	Execution Command (Bigram)
Commander	little to the left	Sensing	Sensing (Unigram)
Commander	there you go	Misc	
Commander	alright you’re on it	Sensing	
Gunner	target	Alert	Sensing (Unigram)
Commander	cease fire cease fire	Termination Command	Termination Command (Bigram)

Preliminary results from these approaches have offered valuable insights into the classification of fire commands, laying the groundwork for further development and refinement in this area. The crew gunnery analysis capabilities are also included in the latest version of the TCAT software.

## **INTEGRATING TCAT WITH GIFT**

---

To allow TCAT and GIFT to communicate with one another to perform automated assessments using NLP capabilities, a number of changes have been made to GIFT's authoring interfaces and real-time assessment engine. Specifically, the team developed TCATInterface, an interoperability interface in GIFT's Gateway module that hosts two-way TCP socket connections between GIFT and TCAT. To support this feature, a message serialization protocol using Google Protocol Buffers was established; the protocol allows GIFT and TCAT to exchange information as binary data over the TCP connections. Next, GIFT's authoring tools were updated to allow TCAT to be used as a training application to perform assessments in GIFT's real-time assessment module. To facilitate these assessments, a new SpeakPhrase condition class was added to GIFT's Domain Knowledge File (DKF) and incorporated with an editor function within the authoring tool interface. This condition leverages the TCP connection to incorporate NLP results from TCAT, specifically text transcripts of spoken language, into assessments. It performs pattern matching on these results to confirm if a specific trainee or learner uttered a required phrase. The required pattern can be defined using a sequence of keywords or a regular expression. To aid in speaker identification during assessments, the TCATInterface's configuration was modified to link unique identifiers, called TCAT instance IDs, to speakers within TCAT. These IDs are assignable to learner-playable team roles within the DKF. Authors using the SpeakPhrase condition editor can specify the intended speaker for a given pattern. When the SpeakPhrase condition initiates assessments, it sends a message to TCAT specifying the speakers being assessed, utilizing these IDs in combination with the TCP sockets.

To demonstrate these changes, a test course was developed and tested alongside the TCATInterface. The goal was to assess whether a specific condition could be used to evaluate gunnery engagement tasks involving multiple speakers. In this setup, TCATInterface was configured with two instance IDs, "Gunner" and "Commander", which were assigned to their respective positions in the DKF. A condition based on the SpeakPhrase model was added to assess whether the gunner spoke the phrase "On the way" before firing. During testing, the condition ignored speech from the commander. However, it registered an "At Expectation" assessment for the gunner only when they uttered the phrase "On the way". Any other phrase resulted in a "Below Expectation" assessment. Initial testing was conducted using the TCAT Example Connector, a Python application that simulates TCAT's behavior through a command-line interface. The communication features of this application are being integrated into TCAT to create a complete pipeline for automated assessments in GIFT, leveraging TCAT's NLP capabilities.

## **CONCLUSIONS**

---

Expanding GIFT's team communication analysis capabilities is essential for facilitating adaptive training in synthetic training environments. TCAT's T5-based model will allow for more accurate dialogue act predictions, eliminate the need for additional preprocessing steps required in previous modeling approaches, and extend TCAT to additionally support information flow classifications. The inclusion of the information flow labels will further enrich investigation of salient team communication patterns, which could provide additional insights into team coordination activities and performance.

As we move forward, we will continue to explore how advances in NLP capabilities can be leveraged to support team communication analytics. The team is currently investigating how large language models (LLMs) can be used to support dialogue analysis. Our approach involves prompt engineering, which

provides models with contextual information regarding the training environment and dataset. A key research objective is to explore prompt engineering strategies leveraging Generative Pre-trained Transformer (GPT) models for inferring dialogue act and information flow labels. This methodology enables the use of limited labeled data instances for training (few-shot learning) and even a description of the task with no labeled data instances (zero-shot learning), rather than requiring large amounts of data for training as in prior approaches. We plan to also explore how prompt engineering can augment the precision of LLMs. In addition, we intend to refine GIFT's condition class logic and authoring interface to include additional communication assessment features, such as dialogue act and information flow matching and speaker interaction patterns. This will allow GIFT to support more robust communication analytics. Finally, we will continue to enhance the generalizability of TCAT's dialogue modeling capabilities to enable team communication analytics for new domains.

## ACKNOWLEDGEMENTS

---

The research described herein has been sponsored by the U.S. Army DEVCOM, Soldier Center under cooperative agreement W912CG-19-2-0001. 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

---

- Jensen, E. L., Pugh, S., & K. D'Mello, S. (2021). A deep transfer learning approach to modeling teacher discourse in the classroom. In M. Scheffel, N. Dowell, S. Joksimovic, & G. Siemens (Eds.), *Proceedings of the LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 302–312). Association for Computing Machinery.
- Johnston, J. H., Phillips, H. L., Milham, L. M., Riddle, D. L., Townsend, L. N., DeCostanza, A. H., ... & Fitzhugh, S. M. (2019). A team training field research study: extending a theory of team development. *Frontiers in Psychology, 10*, 1480.
- Marlow, S. L., Lacerenza, C. N., Paoletti, J., Burke, C. S., & Salas, E. (2018). Does team communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance. *Organizational behavior and human decision processes, 144*, 145-170.
- Min, W., Spain, R., Saville, J. D., Mott, B., Brawner, K., Johnston, J., & Lester, J. (2021, June). Multidimensional team communication modeling for adaptive team training: A hybrid deep learning and graphical modeling framework. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Lecture notes in computer science: Vol. 12748. Artificial intelligence in education* (pp. 293–305). Springer, Cham.
- Pande, J., Min, W., Spain, R. D., Saville, J. D., & Lester, J. (2023). Robust team communication analytics with transformer-based dialogue modeling. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Lecture notes in computer science: Vol. 13916. Artificial intelligence in education* (pp. 639–650). Springer, Cham.
- Pande, J., Paul, S., Min, W., Spain, R., & Lester, J. C. (2023, May). Improving Dialogue Classification Models to Support Team Communication Analytics in GIFT. In A. M. Sinatra (Ed.), *Proceedings of the Eleventh Annual Gift Users Symposium (GIFTSym11)* (pp. 127–136). US Army Combat Capabilities Development Command–Soldier Center.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research, 21*(140), 1-67.
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Spain, R., Min, W., Saville, J. D., Emerson, A., Pande, J., Brawner, K., & Lester, J. C. (2022, May). Leveraging Advances in Natural Language Processing to Support Team Communication Analytics in GIFT. In A. M. Sinatra (Ed.). *Proceedings of the Tenth Annual Gift Users Symposium (GIFTSym10)* (pp. 147–156). US Army Combat Capabilities Development Command–Soldier Center.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

## ABOUT THE AUTHORS

---

**Randall Spain, Ph.D.** is a Research Scientist in the Training and Simulation Division at the U.S. Army DEVCOM, Soldier Center. He holds a Ph.D. in Human Factors Psychology from Old Dominion University. His research focuses on designing and investigating adaptive and intelligent training systems.

**Wookhee Min, Ph.D.** is a Research Scientist in the Center for Educational Informatics at North Carolina State University. He holds a Ph.D. in Computer Science from North Carolina State University. His research focuses on artificial intelligence centering on user modeling, natural language processing, educational data mining, and multimodal learning analytics.

**Nicholas Roberts** is a Senior Software Engineer at Dignitas Technologies and the engineering lead for the GIFT project. Nick has been involved in the engineering of GIFT and supported collaboration and research with the intelligent tutoring system (ITS) community for nearly 10 years. Nicholas contributes to the GIFT community by maintaining the GIFT portal ([www.GIFTTutoring.org](http://www.GIFTTutoring.org)) and GIFT Cloud ([cloud.gifttutoring.org](http://cloud.gifttutoring.org)), supporting conferences such as the GIFT Symposium, and technical exchanges with Soldier Center and their contractors.

**Vikram Kumaran, Ph.D.** is a Research Software Engineer at the Center for Educational Informatics, North Carolina State University. He holds a Ph.D. in Computer Science from the same university. His research is on natural language processing and the application of large language models in generative AI.

**Jay Pande** is a Ph.D. student in Computer Science at North Carolina State University. He received a BS in Computer Science from Duke University in 2020. His research interests lie in the use of speech data to improve educational outcomes for users of adaptive learning environments.

**James Lester, Ph.D.** is the Goodnight Distinguished University Professor in Artificial Intelligence and Machine Learning 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. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI).

# Large Language Models and Their Implications for Conversational Tutors and GIFT

Vasile Rus  
The University of Memphis

## INTRODUCTION

---

Generative Artificial Intelligence (Gen-AI) and in particular Large Language Models (LLMs; Gozalo-Brizuela & Garrido-Merchan, 2023) such as ChatGPT by OpenAI have taken the world by storm in the last couple of years. While Gen-AI's potential across many domains seems to be transformative at first sight, its true potential in various domains is yet to be confirmed and fully understood. The impact on education seems exciting and unsettling at the same time. For instance, as we claimed in Rus (2023), comprehension becomes even more important now in the Gen-AI era as users must comprehend (in order to assess) what Gen-AI tools generate. That is, there is a major shift from generation processes to comprehension processes when using Gen-AI tools. The generation shifts to generating the prompt, i.e., in generating the request describing what you would like to be generated by the Gen-AI tool.

The focus of this chapter is on the potential role of LLMs in developing intelligent tutoring systems (ITSs). Briefly, LLMs are language models which generate output by randomly predicting the next word/idea/item in a sequence according to some learned model, or more simply put, according to some distribution inferred during training. It should be noted that a standard or classical language model (LM) in Natural Language Processing (NLP) is the task of predicting the next word in a natural language sequence of words given the previous context, i.e., the previous words. Importantly, LLMs and more broadly Gen-AI tools are capable of generating other types of content, not only text. That is, LLMs can generate content that includes images (see DALL-E by OpenAI), videos (see Phenaki by Google Research or SORA by OpenAI), audio (see Udio and Suno), and other types of data in response to an input prompt or request, typically, in natural language. Furthermore, LLMs have been shown to offer state-of-the-art solutions to many tasks such as question answering and semantic similarity which, for instance, is used in assessment of student free responses as we will illustrate later.

Our main claim in this paper is that LLMs have the potential to significantly alleviate the authoring bottleneck in developing ITSs which is what the Generalized Intelligent Framework for Tutoring (GIFT; Sottolare et al., 2012) aims to solve. Therefore, a marriage between the two seems worth exploring. Our claim is based on existing evidence that LLMs can generate (multi-modal) content relevant to a given input prompt which, for instance, can specify a learning objective. Furthermore, as already noted, LLMs can help address other tasks such as assessment of student responses. The implication for ITS development is that LLMs can generate learning content and help develop various items and components of ITSs. Importantly, this generation of content and development of other items and components would rely on (mostly) natural language prompts provided by ITS developers. That is, developers can simply provide specifications in natural language in the form of input prompts to LLMs. This is potentially another major advantage as authors, developers, and other stakeholders can specify in natural language what they want as opposed to learning new abstract languages or specialized tools in order to articulate their needs.

We show, based on our experience so far, that LLMs can author content and other items needed for ITS development. While we have not explicitly evaluated the gains in terms of cost and speed of development when using LLMs relative to existing methods for developing the same items, our general sense is that using LLMs will lead to important gains, thus, substantially alleviating but not eliminating (or completely

solving) the ITS authoring bottleneck. As a quick example, we have explored the potential of LLMs with helping us develop a conversational ITS based on the instructional strategy of scaffolded self-explanation for code comprehension (see later for more details). We as well as others have shown that LLMs can automatically generate high quality code examples and accompanying explanations of those code examples which can act as benchmark explanations to assess students' explanations of the code (Oli et al., 2023, 2024). The benchmark explanations are needed to assess students' mastery level (the learner model) as well as inform the next tutorial move, e.g., what feedback and next scaffolding move may be needed (the pedagogical model). Furthermore, LLMs can help automatically assess the correctness of students' explanations, not directly but indirectly by prompting the LLMs to semantically compare the student explanation to the corresponding benchmark explanation (see later). Benchmark explanations have been typically developed manually by human experts which was tedious, expensive, and a major scalability bottleneck. LLMs can generate these benchmark explanations with very high accuracy although not perfectly which means the LLM-generated explanations must be checked by human experts. Furthermore, the generated code explanations do not follow particular theories, e.g., code comprehension theories which was needed in our case (Oli et al., 2023). The advantage is that experts only need to spend a fraction of prior effort for generating explanations as now they only have to check and correct, if needed, LLM-generated explanations.

Indeed, LLMs will not completely solve the authoring bottleneck as they come with new challenges such as hallucinations, i.e., incorrect, made-up parts. That is, when using LLMs, ITS developers, authors, and other stakeholders need to address two new tasks, prompt engineering and LLMs output curation. They will have to acquire new skills and develop new processes that incorporate LLMs for authoring ITSs. These new skills and processes may be discussed by the GIFT community of practice and other relevant communities. For instance, as already noted, a profound implication of the emergence of Gen-AI tools is that reading comprehension becomes even more important (Rus, 2023). Before the launch of ChatGPT and similar LLMs, natural language generation was considered harder than natural language understanding. Yorick Wilks (Dale et al., 1998) famously pointed out that “while the problem of natural language understanding is somewhat like counting from one to infinity, researchers in natural language generation face the problem of counting from infinity to one.” Writing (an essay or computer program) is considered harder than reading (of text or code). Generating language for an essay or generating utterances in a foreign language comes with much more training/practice relative to reading the essay or understanding utterances in the foreign language. In the LLM era, this view/fact flipped which means understanding is now ever more important and will/should be the focus in order to understand and assess the LLMs' output.

To give another example, in Computer Science (CS) education code writing has been emphasized as opposed to code comprehension. We have been claiming, even before the Gen-AI era, that comprehension should be emphasized in CS education as learners and professionals spend a significant portion of their time reading code. To this end, we have been developing code comprehension tutors such as our IES-funded project iCODE: Adaptive Training of Students' Code Comprehension Processes and the NSF-funded project CSEdPad: Investigating and Scaffolding Students' Mental Models during Computer Programming Tasks to Improve Learning, Engagement, and Retention. Indeed, reading comprehension training is becoming very important as LLMs will do much of the draft generation for learners and professionals who must read in order to comprehend and assess the LLM-generated content.

In sum, LLMs are impressive technologies relative to what has been done before in terms of content generation. However, there are several serious challenges that LLMs bring along such as non-deterministic behavior, generating incorrect output or so-called hallucinogenic behavior (generating untrue facts, e.g., person X won award Y in year Z when in fact person W won that award), and data contamination (test data being very similar or identical to training data which can be considered some form of memorization). For instance, many basic code examples can be found in textbooks and on many websites. LLMs have



presumably been trained on such open sources and therefore it is not clear to what extent they simply reproduce, with some variations, what they have already seen or they truly generate totally new, unseen content. Those challenges suggest a more cautious approach to recommending or using such tools too soon without proper, solid, systematic studies that can document the strengths and weaknesses of LLMs. It should be noted that some LLM made-up parts are correct, which makes generative AI positively creative. However, LLMs are not aware of it and thus it requires human effort to assess the true hallucinations from the positive hallucinations, i.e., the truly creative output.

Despite all these challenges, LLMs seem to offer a lot of potential for ITS development and many, if not all, research groups and other interested parties have been exploring the role of LLMs for ITS development. Based on our review of the literature, there seem to be three general approaches to developing ITSs with the help of LLMs. It is beyond the scope of this chapter to fully discuss the pros and cons of these approaches. We will just make some brief comments about each.

A first approach uses prompts to ask an LLM to act like a tutor/instructor. For this first approach, one can imagine two sub-approaches. First, a one-page full-specification of the AI-Tutor can be provided as a prompt as suggested by Open-AI on their website. See *Appendix A* for the full prompt provided by OpenAI with respect to how to build an AI tutor using their LLM ChatGPT. This one-page prompt basically instructs the AI Tutor how to drive the interaction with the student. A second prompt example provided by OpenAI is about building a Socratic Tutor<sup>1</sup>. In both these cases, the LLM plays the more active role, driving the interaction with the learner. In a second sub-approach, one can instruct the learner to first ask ChatGPT what they want to learn about or what problem they need to solve while at the same time briefly instructing the LLM to act like a tutor (a simple prompt specifying the tutor role followed by the instructional need). In this case, the student may need training or external support with respect to prompting the LLM and understanding the LLM output, assessing it, and asking follow-up questions. That is, the learner will play a more active role in guiding the interaction and will need training and/or help on meta-cognitive skills, comprehension, and question asking. It would be interesting to assess these two sub-approaches with respect to their user-friendliness and effectiveness regarding helping students learn as well as with respect to how much training or additional, external support the learner may need in order to interact with the LLM effectively.

This first approach category and in particular the first sub-approach in this category is appealing as most of the development is shifted to the LLM hoping that the right prompt and some minimal or no output curation is needed. The developer of the ITS in this case just focuses on writing the one-page prompt, more or less. In our view, this is a riskier approach with respect to developing highly adaptive, student-tailored ITSs as there is plenty of evidence that prompt engineering is still more of an art than science and the LLMs' output contains hallucinations. We claimed in Rus (2023) that detecting hallucinations in LLMs' output is the next killer app. Until this is solved, this approach is too risky, particularly, for well-defined, rigorous domains such as programming or physics (as opposed to more open-ended domains such as essay writing or debating).

The second approach relies on using LLMs to emulate an existing tutoring framework and given that existing LLMs' main focus is on generating language this approach works well for conversational ITSs. A somehow simplistic example of this approach is the approach proposed recently by Schmucker and colleagues (2023). They present a process using LLMs to automatically induct conversational tutoring systems based on the Expectation-Misconception tailored (EMT) dialogue framework used in AutoTutor (Graesser et al., 2004). While this approach is trying to emulate the EMT framework using LLMs, the result is just a simplistic (naive) conversational companion for a textbook or other sources of 'lesson content' that form the input to their approach. The proposed process is naive and we would not call the result an ITS for

---

<sup>1</sup> <https://platform.openai.com/examples/default-socratic-tutor> (retrieved on April 14, 2024)

many reasons. For instance, there is no student model that can reflect at each moment students' level of mastery. The instructional tasks are just verification questions about the content of an e-book as opposed to more complex tasks such as problems to be solved by the learner. It is not clear what the nature of these verification questions is either. Are they shallow, factoid questions (*what?*) or deeper, more conceptual questions (*why?*). As yet another shortcoming of their approach, there are no misconceptions identified or documented, instead the authors simply indicate they rely on 'GPT4's ability to detect factually incorrect information', thus limiting the number of learning opportunities such as correcting immediately well know misconceptions the learner may articulate. This overreliance on GPT4's abilities seems unfounded as we have shown recently (Banjade, Oli, & Rus, in press). Indeed, we have evidence that LLMs are not very good at identifying missing parts or incorrect parts in student generated responses when prompted directly to do so (Banjade, Oli, & Rus, in press). They do a much better job when the task is reframed as a semantic similarity task for which the correct and incorrect (misconceptions) are readily available.

A third approach, which is the one we emphasize here, employs LLMs for specific tasks needed to develop ITSs such as inferring a domain model or generating the next hint. This third approach explores the role of LLMs on specific ITS-related tasks as we will illustrate throughout this chapter. The advantage is that task specific prompts and output curation strategies can be designed which should lead to higher quality, less risky adaptive training systems. The focus is to use LLMs to speed up and reduce development costs as opposed to replacing entirely existing ITS architectures and relying entirely on LLMs. Furthermore, this more task specific approach will enable us to better understand the pros and cons of using LLMs for ITSs which is much needed at this early, hype stage of Gen-AI. Furthermore, we argue for ITS developers to build or augment their existing ITS platforms with LLM-based components only after clear evidence that LLMs can lead to improvements and that well understood and documented processes with respect to how to use LLMs in ITS development are in place. In this approach, existing ITS platforms will drive the interaction with the learners based on decades of development, testing, and refinements while making use of LLMs where clear evidence justifies such use.

### **Evidence LLMs Can Alleviate The Authoring Bottleneck In ITS Development**

As pointed out earlier, LLMs/Gen-AI can generate text, images, videos, and other content, e.g., code examples. Additionally, they have been shown to provide state-of-the-art solutions to many natural language processing tasks such as semantic similarity which is of key importance to understanding what students say and therefore assessing their mastery level (part of the learner model) in conversational ITSs. More generally, we make the case that LLMs can be used to develop all major components of ITSs such as content creation (e.g., instructional items that can serve a particular learning goal such as code examples meant to offer learners the opportunity to master a particular programming concept), domain models, student models, as well as pedagogical models. Furthermore, LLMs can be used to author the interaction/interface model but we will not address this component here. We make our case in the context of dialogue-based ITSs as most of our work is in the context of such ITSs. Furthermore, LLMs can best fit such ITSs as LLMs' main focus is text-to-text generation. Multi-modal input and output, i.e. language mixed with other type of content such as images is possible but language is the primary output of existing LLMs.

Based on our recent explorations of using LLMs for conversational ITS development, we provide next evidence that LLMs can help author domain models and related instructional items linked to the knowledge components identified by the *domain model*, to develop assessment components which are critical to inferring learners' mastery level and other states, e.g., emotional state, which are part of the *learner model*, and to generate hints/questions which are components of instructional strategies which in turn are part of the *pedagogical model* of ITSs.

### *LLM-based Authoring of Domain Models*

Our team has recently explored the potential of LLMs to author domain models. We needed a domain model and corresponding instructional items in the form of code examples to alleviate authoring costs associated with developing the iCODE project, an ITS for adaptive training of code comprehension skills. To the best of our knowledge, no widely available domain model exists for the intro-to-programming domain.

Specifically, we explored how LLMs can extract domain models from textbooks, e.g., intro-to-programming textbooks. We used LLMs to address the following two tasks separately: concept or knowledge component (KC) extraction and relationship extraction. For the KC extraction task, we used only zero-shot and one-shot prompting due to limitations on the size of input one can provide to LLMs as we needed to provide the prompt and the text, e.g., a textbook section, from where to extract the KCs. Results so far indicate LLMs can extract KCs from textbook texts with precision in the 60s-70s% and recall in the range of 20s-70s% as judged at rank 5, 10, and 15, i.e., when considering the top 5, 10, and 15 terms extracted. Extracting general relations between KCs has not led to any promising results which is not surprising as relation extraction is harder than concept extraction, a well-known fact in the field of information extraction. A more specific relation extraction task in which we only focused on 4 relations (is-a, part-of, has-property, and none) has led to much better results with precision and recall in the 80s-90s%.

It should be noted that, we also tried a simple prompt, “Generate a domain model for the intro-to-programming domain. The domain model should include a list of key computer programming concepts that students learning to program need to master”. The response was pretty good but not detailed enough for ITSs which handle both macro- and micro-adaptation. The generated list of concepts could be useful, for instance, for open learner models, i.e., to provide students and instructors and other stakeholders a summary of students’ performance, a summary which is not too coarse and not too detailed but good enough for human use.

### *LLM-based Authoring of Instructional Items*

We have also explored the potential of LLMs to author instructional items in the form of code examples and natural language explanations of those code examples. Automating the task of creating code examples and accompanying explanations aligned with the concepts in the domain model could lead to substantial advantages in terms of scaling up the use of explanations across topics and domains.

We have shown that LLMs can generate the kind of examples we need, i.e., code examples used in intro-to-programming courses (CS1 and CS2). We also explored how well LLMs can generate explanations. The correctness of the explanations of the code examples was high (90+%) whereas the completeness was good too (80+%). Nevertheless, the kind of explanations generated was not of the kind we wanted. That is the case as we developed our own, theory-driven explanations (Rus, 2022) which LLMs were not trained to generate. We do plan to explore few-shot prompting, providing few examples of the kind of explanations we want, and assess to what degree LLMs can generate such theory-driven explanations.

To provide more details, we studied the behavior of LLMs with respect to the task of code explanation generation and more specifically how five major input parameters alter the output of the LLMs, i.e., how the generated code explanations vary while those input parameters vary. We focused on the following five input parameters: input prompt wording, code example type, temperature, LLM model, and programming language. All LLMs we tried generate widely different types of explanations depending on the actual wording of the input prompt, temperature parameter, and code example type. Overall, all LLMs showed a great deal of diversity, which can be seen as a form of inconsistency as differences in the nature and characteristics generated explanations, e.g., length and readability level, are substantial. There are three major consequences of this diversity/inconsistency: (1) the exact parameters used by researchers to generate

code explanations must be well documented as otherwise their work cannot be used or replicated; (2) it is challenging to use LLMs' diverse/inconsistent output for certain pedagogical needs that are supposed to rely on explanations of code that consistently follow a particular theory without additional work; and (3) LLMs may be used to obtain a set of rough explanations after which a substantial human effort would be needed to refine those explanations.

A comprehensive understanding of how various factors influence LLM behavior in code explanation, including prompt creation, is essential to establish clear guidelines for using these technologies in education. Such guidelines are needed for other uses of LLMs for ITS development. To facilitate the use of LLMs in generating code explanations, a repository of prompts with user ratings and parameters like temperature should be established. This may be necessary for all uses of LLMs in ITS development and thus the GIFT community should discuss and create prompting guidelines and repositories of prompts.

### *Automated Assessment of Student Responses*

In the context of conversational ITSs, the main form of interaction between the learner and the system is natural language. Assessing students' natural language statements is a major challenge in such systems while at the same time a critical component as many other components depend on it such as the learner model, the pedagogical model, and the interaction/interface model. We have explored how to use LLMs to assess students' natural responses in two separate experiments.

In a first experiment, we explored various prompting strategies and modelled the student answer assessment task as a semantic similarity task. Four different large language models were used: OpenAI's ChatGPT-3.5-turbo-0613, ChatGPT-4-0613, gpt-4-1106-preview (GPT-4 Turbo), and Meta's open source model LLaMa2-chat2. We opted for deterministic results by setting the temperature parameter to 0 and set a maximum token length of 1200 to limit the scope of the generated sequences.

In terms of prompting, we used simple prompts asking the LLMs to compute the semantic similarity between student responses and benchmark responses provided by experts. The prompt instructed the LLM to predict the similarity score on a scale of 1-5, mirroring human judgment, with 1 indicating no semantic similarity and 5 indicating semantic equivalence between the pair of sentences. Alternatively, we have prompted the LLMs to predict a normalized similarity score within the range [0,1], leveraging large language models' strong textual reasoning and their exposure to percentage-related data during pre-training. In addition, we also explore advanced prompting strategies. These include the conventional few-shot prompting, also known as in-context learning, where the LLM is tasked to infer from provided examples or task descriptions, as well as few-shot chain-of-thought (CoT) prompting where the LLM is guided to perform a task step by step.

There are three main findings of this experiment: (i) GPT-4 offers best performance (Spearman's correlation of 0.82 with human scores) on par with specialized models, e.g, RoBERTa Sentence Transformer; (ii) prompting the LLMs to predict semantic similarity on a scale of 0-1 yields superior performance (Spearman's correlation of 0.82) compared to prompting to predict similarity on other scales (1-5; Spearman's correlation of 0.70), and (iii) the advance strategies consistently boost ChatGPT's performance, with manual chain-of-thought (CoT) providing the most significant benefits. Notably, the standard few-shot CoT enhances Chat-GPT's overall performance (on average 15% better than baseline prompting for ChatGPT based model) with GPT-4 providing the best performance for our task.

In a second experiment, we investigated two LLM-based methods based to identify gaps or missing parts in learners' self-explanations. This is an important task as, for instance, a typical next tutorial move is about asking/hinting learners to think about the missing parts in their solution to a problem or explanation of a code examples. Before generating the hint, one must identify what is missing in the student response. As

noted, we experimented with two methods and four distinct LLMs (GPT-3.5, GPT-4, LLAMA2, and MIXTRAL) in two distinct settings, simulated versus actual student data. The two approaches were (1) a holistic approach in which LLMs are prompted to identify the missing parts given a student explanation and the corresponding code example and (2) a point-wise, semantic similarity approach at sentence level in which student explanations are broken down into units of analysis (sentences) and then LLMs are prompted to identify which such sentences match corresponding sentences in the benchmark explanations (a sentence in the benchmark/reference explanation is called an expectation). Results revealed that zero-shot prompting for explanation gap identification (the holistic method) yields suboptimal results, whereas the semantic similarity method significantly improves task performance. As before, GPT-4 leads to best results with an accuracy of 94% for the pointwise method compared to the best holistic method result which was for GPT-3.5 with an accuracy of 48%.

The immediate lessons from these two experiments are threefold. First, the exact modelling of a target task when using LLM-based methods is very important (see the scoring scale in the first experiment/task and the holistic versus pointwise modelling in the second experiment/task). Second, there is a wide range in performance across various LLMs with GPT-4 leading to best results in most cases. Third, LLMs perform comparably well, in particular GPT models, to fine-tuned encoder-based models. This is in general true about ChatGPT, i.e., it provides state-of-the-art solutions on many tasks on par with more fine-tuned, dedicated solutions to these tasks. That is, it is a general model that works as well as more specialized solutions on many tasks – a Swiss knife kind of tool. While it does not completely solve many of these tasks it provides state-of-the-art solutions.

### *Hint Generation with LLMs*

Finally, we are in the process of exploring how LLMs can generate the next hint in the context of a conversational tutor for code comprehension that implements scaffolded self-explanation as the main pedagogical strategy. We prompt LLMs to generate the next hint given the dialogue history which includes the student performance so far and the target content the student is supposed to master. For instance, as described earlier, one can identify what is missing in a student response and use that as the target content based on which to generate the next hint for the next tutorial move. Results so far are promising in terms of alleviating the assessment of students' overall mastery and specific responses and authoring of hints.

## **CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH**

We have made the case that LLMs can potentially alleviate the ITS authoring bottleneck. While early evidence is promising, the true potential of LLMs is yet to be confirmed and fully understood. We provided evidence based on a number of experiments with LLMs to address various tasks such as domain modelling, auto-assessment of freely generated natural language responses, identifying gaps in learners' explanations, and hint generation. These tasks span three of the four major ITS modules specified by the GIFT architecture: domain model, learner model, and pedagogical model. Our results indicate that LLMs can generate various items and components of ITSs which humans must eventually curate and refine. While there is a perceived substantial boost in scalability of ITS development as developers do not have to start from scratch, the exact level of increase in scalability is yet to be fully understood and quantified. LLMs bring along their own challenges and shortcomings. For instance, as noted earlier, they cannot, by default, generate code explanations that follow our types of explanations designed based on code comprehension and learning theories. By extension, it is yet to be found if LLMs can generate other types of ITS items based on other relevant theories and, if not, what potential solutions there may be. One reviewer of this paper wondered what we think of using LLMs to drive an after action review (AAR) after a learning event

concluded. The answer is yes, it is possible, in theory. How well the LLMs can do it at this moment should be explored empirically. For instance, the LLM can be used to compare the expected outcomes with the actual outcomes. Any gaps or discrepancies can be used to drive a conversation as illustrated earlier for the task of identifying gaps in student explanations of code examples. If the performance outcomes are very specific and also numerical, we believe the task becomes more challenging but worth exploring.

Another big question, also asked by a reviewer, is whether LLMs can replace conversational tutors' or other ITSs functionality when error and hallucinations are effectively addressed. In theory, LLMs should not be able to replace existing ITS platforms as there are too many shortcomings of approaches that emulate ITSs with LLMs. For instance, existing LLMs do not have an explicit learner model and they are well known to not do very well when it comes to quantitative reasoning which is needed for estimating student performance. This could be a good empirical question and we would be curious to be proven wrong in our prediction.

Finally, we were asked about using LLMs to drive development of immersive learning scenarios. Again, in theory, LLMs could be used to assist with the development of immersive learning scenarios. Assuming specifications of such scenarios can be done in some kind of mix of natural and artificial language, then it is indeed theoretically possible. How soon this could be possible in practice is a good empirical question.

Given that LLMs come with new tasks and challenges, there is a need to re-skill and re-think processes when using LLMs for ITS development, i.e., prompt engineering and comprehension skills take a more crucial role that the GIFT community of practice must discuss and strategically address. Prompting guidelines, prompt repositories, and reproducibility of work when using LLMs are critical aspects that the GIFT community must address as well. To conclude, LLMs are an interesting and promising new technology which everyone is inspired and excited about. After the hype is over, a more realistic view and applications of LLMs will emerge.

## ACKNOWLEDGEMENTS

---

This research was sponsored by DEVCOM SC-TSD-STTC and was accomplished under Cooperative Agreement Number W912CG-24-2-0001. 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 DEVCOM SC-TSD-STTC or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein. This work has also been partially supported by the following grants awarded to Dr. Vasile Rus: the Learner Data Institute (NSF award 1934745); CEdPad (NSF award 1822816); and iCODE (IES award R305A220385). The opinions, findings, and results are solely those of the authors and do not reflect those of NSF and IES, or DoD. Neither NSF or IES have approved or endorsed its content.

## REFERENCES

---

- Banjade, R., Oli, P., & Rus, V. (in press) Identifying Gaps In Students' Explanations of Code Using LLMs, To appear in Proceedings of 25th International Conference on Artificial Intelligence in Education, July 8-12, 2024, Recife, Brazil
- Dale, R., DiEugenio, B., and Scott, D. (1998). Introduction to the Special Issue on Natural Language Generation. *Computational Linguistics*, 24(3), 345–353.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Graesser, A. & Lu, S. & Jackson, G. & Mitchell, H. & Ventura, M. & Olney, A. & Louwerse, M. (2004). AutoTutor: A Tutor with Dialogue in Natural Language. *Behavior Research Methods*, 36, 180-192. 10.3758/BF03195563.
- Gozalo-Brizuela, R. & Garrido-Merchan, E. (2023). Chatgpt is not all you need. A state of the art review of large generative ai models. arXiv preprint arXiv:2301.04655, 2023.
- Oli, P., Banjade, R., Tamang, L.J., Rus, V. (2023). The Behavior of Large Language Models When Prompted to Generate Code Explanations, NeurIPS'23 Workshop on Generative AI for Education (GAIED), New Orleans, 15 December, 2023.
- Oli, P., Banjade, R., Chapagain, J., and Rus, V. (2024) *Automated Assessment of Students' Code Comprehension using LLMs*, Proceeding of AAAI 2024 Workshop on AI for Education - Bridging Innovation and Responsibility, February 26-27, 2024, Vancouver, Canada.
- Rus, V., Brusilovsky, P., Tamang, L.J., Akhuseyinoglu, K., & Fleming, S. (2022). DeepCode: An Annotated Set of Instructional Code Examples to Foster Deep Code Comprehension and Learning. Proceedings of 18th International Conference on Intelligent Tutoring Systems. Retrieved from <https://par.nsf.gov/biblio/10367910>.
- Rus, V. (2023). Generative AI And Its Impact: An Educator and AIED Researcher's View. Accessed Online on March 25, 2024. <https://drive.google.com/file/d/1SCwkqViflItHHNvnpQQIWhzFQ19xZyjq/view?usp=sharing>
- Schmucker, R., Xia, M., Azaria, A., and Mitchell, T. (2023). Ruffle and Riley: Towards the Automated Induction of Conversational Tutoring Systems, NeurIPS 2023 Workshop on Generative AI for Education (GAIED), December, 2023
- Sottolare, R. & Brawner, K. & Goldberg, B. & Holden, H. (2012). The Generalized Intelligent Framework for Tutoring (GIFT). 10.13140/2.1.1629.6003.

## APPENDIX-A

---

*OpenAI suggested prompt to develop AI tutors with their ChatGPT LLM (Retrieved on April 14, 2024, from <https://openai.com/blog/teaching-with-ai>):*

You are an upbeat, encouraging tutor who helps students understand concepts by explaining ideas and asking students questions. Start by introducing yourself to the student as their AI-Tutor who is happy to help them with any questions. Only ask one question at a time.

First, ask them what they would like to learn about. Wait for the response. Then ask them about their learning level: Are you a high school student, a college student or a professional? Wait for their response. Then ask them what they know already about the topic they have chosen. Wait for a response.

Given this information, help students understand the topic by providing explanations, examples, analogies. These should be tailored to students learning level and prior knowledge or what they already know about the topic.

Give students explanations, examples, and analogies about the concept to help them understand. You should guide students in an open-ended way. Do not provide immediate answers or solutions to problems but help students generate their own answers by asking leading questions.

Ask students to explain their thinking. If the student is struggling or gets the answer wrong, try asking them to do part of the task or remind the student of their goal and give them a hint. If students improve, then praise them and show excitement. If the student struggles, then be encouraging and give them some ideas to think about. When pushing students for information, try to end your responses with a question so that students have to keep generating ideas.

Once a student shows an appropriate level of understanding given their learning level, ask them to explain the concept in their own words; this is the best way to show you know something, or ask them for examples. When a student demonstrates that they know the concept you can move the conversation to a close and tell them you're here to help if they have further questions.

## **ABOUT THE AUTHORS**

---

*Vasile Rus, Ph.D. is a Professor of Computer Science at The University of Memphis where he also heads the Data Science Center and Research Cluster. Dr. Rus' academic career accomplishments are numerous including a professorship, several best paper awards (all with his advisees), and multi-million federally funded grants. Dr. Rus is also a faculty in the Institute for Intelligent Systems where he is collaborating with Cognitive Scientists on developing adaptive educational technologies. Dr. Rus' research interests are human and machine learning whereas his overall research goal is to make the learning ecosystem more effective, equitable, engaging, efficient, relevant, and affordable.*



# Integrating Machine Learning Models and GIFT

Andy Smith<sup>1</sup>, Randall Spain<sup>2</sup>, Nicholas Roberts<sup>3</sup>, Jonathan Rowe<sup>1</sup>, Bradford Mott<sup>1</sup>, and James Lester<sup>1</sup>

<sup>1</sup>North Carolina State University, <sup>2</sup>U.S. Army Combat Capabilities Development Command (DEVCOM) - Soldier Center, Dignitas Technologies<sup>3</sup>

## INTRODUCTION

---

Numerous studies conducted over the last two decades have highlighted the effectiveness of adaptive instructional systems (AISs) for improving learning outcomes (Mousavinasab et al., 2021; Xu et al. 2019; VanLehn, 2011). In addition to their use in academic environments, AISs play a significant role in supporting simulation-based training for the U.S. Military (Johnston et al., 2018; Kulik & Fletcher, 2016). Integrating the functions of an AIS such as automated coaching, feedback, and instructional support into synthetic training environments continues to be an area of active research, particularly for collective training (Smith et al., 2022; Spain et al., 2021). A critical requirement is the creation of data-driven models that can dynamically determine the most suitable time, content, and approach for providing teams with instructional support as they conduct training in highly dynamic synthetic training scenarios (Kochmar et al., 2022; Smith et al., 2023).

The Generalized Intelligent Framework for Tutoring (GIFT), an open-source modular framework for designing and developing AISs, has become a key architecture and platform for testing and investigating next-generation adaptive and intelligent training systems (Roberts et al., 2023). Researchers have used GIFT to train marksmanship fundamentals, land navigation and terrain association skill, dismounted infantry battle drills, course of action planning, and counterinsurgency doctrine, providing learners with adaptive training experiences to enhance training effectiveness. In recent years, GIFT has been enhanced to support adaptive training for teams. These upgrades include incorporating team structure modeling into GIFT's Domain Knowledge File (DKF), adding new condition classes and scenario modifications to support team evaluation, delivering feedback at individual and team levels, and modifying the Game Master Interface to inject scenario modifications during collective simulation-based training events and support after action review.

At last year's annual symposium, we presented enhancements to GIFT that build upon these capabilities with an aim for supporting data-driven models of adaptive team feedback (Smith et al., 2023). These models can guide tutorial planning and feedback using reinforcement learning (RL)-driven models. Using crew gunnery as an example, we showed how we modified GIFT to facilitate the exploration of different coaching strategies. We introduced modifications that provide flexibility in content (feedback covering multiple course concepts) and feedback modality (textual vs. graphical) and allow for varying the timing of feedback, enabling partially delayed and summative feedback. These enhancements were achieved by creating a custom strategy that can pass assessment data from GIFT's DKF to an External Strategy Provider and display feedback from that server within GIFT's Tutor User Interface (TUI) or through a customizable web page. Additionally, we introduced the capability to support micro-adaptive sequencing within Virtual Battlespace 3 (VBS3) scenarios by modeling a VBS3-based training exercise as a series of course objects in GIFT's course creator interface, each with its own DKF. This facilitates the automatic sequencing of individual engagements with GIFT without requiring an operator controller or trainees to restart VBS3.

In this paper, we describe additional enhancements to GIFT that support the goal of providing teams with data-driven coaching and feedback during simulation-based training exercises. Specifically, we describe how we have extended the External Strategy Provider to support machine-learned models so that GIFT can be used to train and deploy policy-driven coaching models using offline and online RL. The motivation for this work stems from previous research designing data-driven tutorial planning policies in GIFT. At that time, we integrated a Markov decision process-based model into GIFT that used custom weights and

configurations to support adaptive remediation (Spain et al., 2019). The disadvantages of this approach are that the pedagogical framework only supported MDP (Markov Decision Process) policies and it was difficult to experiment with different models. Furthermore, the models were fixed, so the policy could not self-improve overtime. The External Strategy Provider addresses these limitations by establishing a corresponding run-time data pipeline for processing data on the states of learners, the training environment, and adaptive tutor to serve as input to the coaching model and translate coaching strategies into pedagogical tactics for delivery to distributed teams of learners. The feedback server framework allows for data-driven models to be easily connected to existing GIFT courses, and for rapid prototyping of different models with minimal changes to the GIFT course.

We discuss future directions for these enhancements, including incorporating feedback information into the user's xAPI record, as well as generalizing the modifications to better serve a wider range of training domains and scenarios. Further we discuss the need to design a new GIFT course feature that will allow researchers to establish linkages between RL-based coaching and macro-adaptive and micro-adaptive pedagogical decision-making features in GIFT. GIFT presently supports two pedagogical workflows: a macro-adaptive course flow model based on Merrill's Component Display Theory (Merrill, 1983) and a micro-adaptive remediation model grounded in the ICAP framework (Chi & Wylie, 2014). A key opportunity lies in creating new interfaces that would allow authors to use the External Strategy Provider to create and deploy RL-based coaching models with both macro-adaptive and micro-adaptive pedagogical decision-making capabilities in GIFT. Addressing these requirements will position GIFT to support generalized data-driven team coaching capabilities, showing great promise in enhancing intelligent tutoring for collective training within simulation-based training environments.

## RESEARCH CONTEXT

---

The U.S. Army Learning Concept 2030-2040 (U.S. Army, 2024) seeks to integrate adaptive training capabilities into virtual training experiences to enhance training effectiveness and team performance. An important task towards the development of adaptive training for teams is operationalizing team-level constructs that should be assessed and remediated during training. Providing team members with coaching and feedback is crucial for supporting the development of team interdependence and adaptability. In particular, feedback and coaching can help teams improve the following team processes:

1. **Interdependence:** Effective teams rely on the contributions of their members to achieve shared goals. Guided feedback and coaching can enable team members to understand their roles and responsibilities and develop effective collaboration strategies.
2. **Shared Mental Models:** Effective teams have a shared understanding of their goals, roles, and processes. Guided feedback and coaching can facilitate the development of shared mental models, allowing team members to anticipate each other's actions and coordinate their efforts seamlessly.
3. **Communication and Information Exchange:** Teams must pass and exchange information with other team members to ensure a collective understanding and awareness of the operating environment. Messages must be clear, concise, and provided in a timely manner. Guided feedback and coaching can help team members develop effective communication skills, learn to provide information constructively, and improve team coordination.
4. **Continuous Improvement:** Effective teams engage in continuous improvement to maintain their success. Guided feedback and coaching can provide teams with the tools and strategies to assess their performance, identify areas for development, and create plans for ongoing growth.

Devising feedback and coaching strategies that can be applied within virtual training experiences to support these constructs poses a unique challenge, as it requires the development of computational models capable of determining when, how, and what type of feedback and support to deliver, taking into account the complex dynamics of team interactions and the need for real-time adjustments to instructional strategies.

In recent years, machine learning techniques, including RL and those based on MDPs, have demonstrated promise as data-driven approaches for modeling pedagogical coaching and feedback in AISs (Fahid et al., 2024; Fahid et al., 2021; Georgila et al., 2019). RL has shown particular potential for inducing tutorial policies that optimize student learning outcomes without the need for manually programmed pedagogical policies or expert tutor demonstrations. Tutorial planners control how scaffolding (e.g., feedback, coaching, hints and support) is structured and delivered to learners to create dynamically personalized learning experiences. To date, most research on RL-based models of tutorial planning has focused on individual learners. For instance, in our prior research we induced data-driven tutorial policies to provide adaptive scaffolding based on the Interactive, Constructive, Active, Passive framework for cognitive engagement using reinforcement learning for individual learners (Spain et al., 2022). Results showed that the best performing policies optimized learning gains by inducing an adaptive fading approach in which learners received less cognitively engaging forms of remediation as they advanced through the training course. This policy was consistent with preliminary analyses that showed constructive remediation became less effective as learners progressed through the training session. Results also showed that learners' prior knowledge impacted the type of scaffold that was recommended, thus showing evidence of an aptitude–treatment interaction.

In this work, we aimed to extend this line of research to deliver coaching and feedback at the team level. In particular we seek to integrate machine learning models into the GIFT architecture to allow for rapid prototyping and deployment in a variety of training domains and contexts.

## **CREW GUNNERY TRAINING COURSE**

---

To develop and design adaptive team feedback models using GIFT, we are using crew gunnery training in Virtual Battlespace 3 (VBS3) as a testbed. These virtual scenarios were developed at the Warrior Skills Training Center (WSTC) in Fort Cavazos, Texas and allow crews to rehearse and practice crew coordination activities before engaging in live-fire qualification exercises. The VBS3 crew gunnery scenario includes a series of six engagements. Each engagement involves detecting, identifying, and engaging stationary or moving targets. During this direct fire engagement process, the coordination of actions and behaviors among the vehicle commander, gunner, and driver is crucial. After identifying a threat, crews engage in a fire command sequence, a defined protocol for communicating information to facilitate a coordinated response. Our crew gunnery assessment model, represented as a DKF in GIFT, includes task and team-specific concepts for each engagement, providing the potential to support different forms of feedback and support at the crew and individual levels to remediate and facilitate reflection to improve crew performance.

## **GIFT MODIFICATIONS TO SUPPORT ML-DRIVEN FEEDBACK**

---

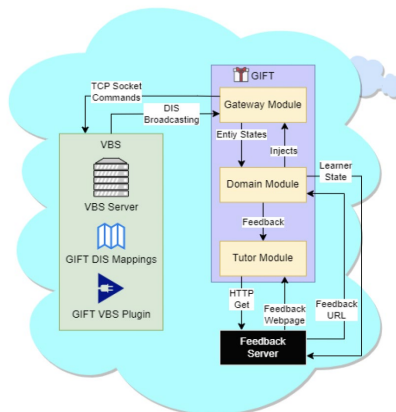
To enable feedback to be driven by Machine Learning (ML) models, we first expanded the set of trainee performance data passed from GIFT. At last year's GIFTSym, we presented an initial version of this protocol, where GIFT was modified to pass assessment information via an HTTP web request (Smith et al., 2023). This protocol has been expanded to include all concept class variables used as part of any automated assessments, and is now included in the main GIFT distribution as part of the External Strategy Provider.

For the crew gunnery domain this consists of several performance metrics used in existing score sheets to evaluate performance, such as time to identify target, time to engage target, and time to eliminate target. Additionally using capabilities afforded by the VBS3 simulation environment, we also detect the time to orient the firing reticle on the target. These performance variables are automatically calculated, and passed to the External Strategy Provider for each target in a given engagement. See Table 1 for assessment variables for the Crew Gunnery course.

**Table 1. Assessment variables for Crew Gunnery Course**

Name	Unit	Description
<i>Target Up</i>	Timestamp	Timestamp each target is raised
<i>Target Detection Time</i>	ms	Time to when target enters gunner field of view from <i>Target Up</i>
<i>Target Orientation Time</i>	ms	Time to when reticle is on target from <i>Target Up</i>
<i>Open Time</i>	Timestamp	Timestamp of when firing begins
<i>Target First Strike</i>	Timestamp	Timestamp when target is first hit
<i>Kill Efficiency</i>	ms	Time to when target is killed from <i>Target First Strike</i>
<i>Kill Time</i>	ms	Time to when target is killed from <i>Target Up</i>
<i>No Kill Time</i>	ms	Time target is lowered if not killed
<i>Target Close Time</i>	Timestamp	Timestamp of when target is lowered
<i>Target Engagement Time</i>	ms	Time last target is killed from first <i>Target Up</i>

The External Strategy Provider protocol was also expanded to include performance data from previous engagements in the training course. For each of the modular course objects in the DKF, all performance and assessment data is passed for that engagement, as well as all prior engagements in the course as part of a *History* object in the JSON passed to the External Strategy Provider (ESP). This allows ESPs to more easily incorporate data from the current scenario, without having to maintain user session information outside of GIFT. A diagram of the architecture is shown below in Figure 1, with the External Strategy Provider labeled as Feedback Server.



**Figure 1. Architecture diagram of VBS3, GIFT, and External Strategy Provider**

## EXTERNAL STRATEGY PROVIDER ENHANCEMENTS TO SUPPORT ML-DRIVEN FEEDBACK

The current External Strategy Provider included in the GIFT distribution demonstrates the capability of providing feedback to trainees in the form of an HTML webpage. Decisions on what webpage to provide to the users are made by changing the feedback URL in GIFT, or by modifying the server to vary which version of feedback is shown based on information such as the role of the user, or even random assignment.

This year, we demonstrate an enhanced version of the ESP to allow for feedback decisions driven by a machine learned model. To accomplish this we first created a separate Python web server to run in parallel to the existing ESP web server. This is mainly for convenience, and future versions will replace the existing ESP server functionality with one Python server to reduce complexity. Python was chosen due to its support of a wide variety of popular machine learning and data analysis toolkits and frameworks, such as Tensorflow and PyTorch for deep learning applications, as well as the SciKit-Learn toolkit which supports implementations of a large number of model architectures for classification, regression, and clustering. The server architecture allows for models to be trained using any of these techniques, as long as the trained model can be saved to a file for later inference. Then, upon receiving performance information from GIFT, the server can perform inference on the trained model and determine which type of feedback to provide to the trainee.

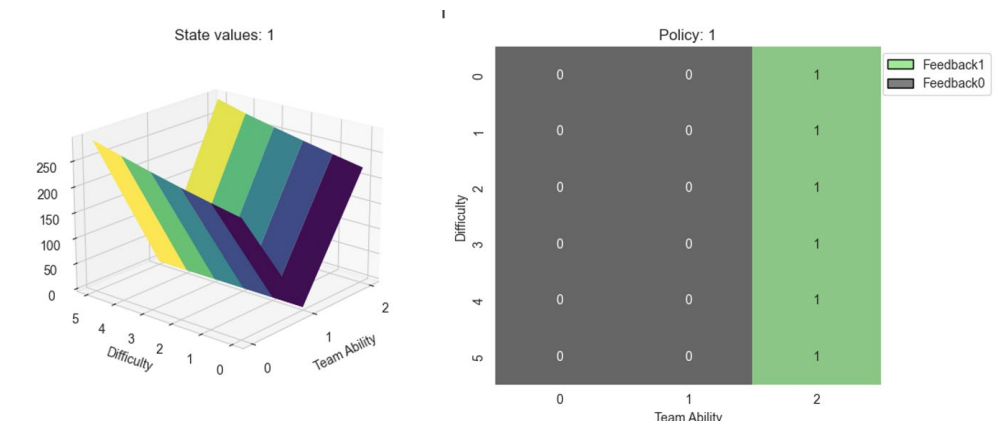


Figure 2. Cumulative reward and q-value table for an example simulation model

To illustrate these enhancements, consider a simple Q-learning model. Q-learning is a model-free reinforcement learning algorithm that learns a value for each possible action, given a particular state. As with most reinforcement learning architectures, this model seeks to learn a policy  $P$ , that determines which action to take from a set of actions  $A$ , given a state representation  $S$ , to maximize reward  $R$ . See Figure 2 for a cumulative reward and q-value table for an example simulation model. In the crew gunnery course, the state representation is passed by GIFT to the server after each engagement. In this example we will only use two dimensions of this data, the Prior Crew Ability Level (0, 1, or 2), and the Engagement Difficulty Level (0 thru 5), which is calculated using the target information (number, type, distance) for the engagement passed from GIFT. The learned policy will determine which action from  $A$  to perform after each engagement, in this example the actions are to provide Feedback Type 0 or Feedback Type 1 to the trainees. A policy was learned by training the model with simulated students, and an underlying simulation model was chosen that boosted reward for beginner (Level 0) trainees when receiving feedback type 0, and boosted reward for experts (Level 2) when receiving feedback type 1. As expected, the model converges on a policy illustrated in the table below, where the model always provides type 0 feedback to Level 0 teams, and feedback type 1 to level 2 teams. Level 1 teams also receive feedback 0 by default since their Q-value is 0 for both types of feedback. This table is then serialized to a file via the *pickle* Python function,

and then deserialized and used for model inference when receiving a request from GIFT. While this model and simulation are overly simplistic, it highlights the new functionality added to the ESP, as well as the flexibility to support inference from multiple types of models.

The framework can also be easily modified to support Online RL, where the model constantly adapts and improves as it receives new data. The main required enhancement from the previous example, is the feedback server must now persist additional information beyond just the current policy. Upon receiving a feedback request from a GIFT client, the server now performs both a model inference and model update step, where the policy weights may be updated if the given step has resulted in a reward. While online models have been shown to have advantages over offline ones for educational applications (Fahid et al., 2024), and are relatively easy to implement in the new setup, to effectively deploy it in an actual training environment does have additional logistical complexities such as all machines using the training course requiring access to the same feedback server, rather than just an instance of the server with the previously trained policy.

## **CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH**

---

Providing effective coaching and feedback to teams is a key challenge for AIs. Utilizing machine learning techniques to drive this feedback shows significant promise for providing data-driven adaptive pedagogy, as well as providing empirical evidence to support existing theoretical models of feedback and coaching. However, implementing and training these models in GIFT is very cumbersome and labor intensive. The new ESP protocol and additional enhancements highlighted in this paper greatly ease this burden, and allow for rapid prototyping of ML-based feedback models for GIFT courses.

In the next year we will focus on improving the generalizability of the framework by applying it to a new domain. Additionally we will focus on incorporating the new enhancements into the main GIFT distribution branch, as well as using the enhancements to investigate data-driven feedback for teams in the domain of crew gunnery.

## **ACKNOWLEDGEMENTS**

---

The research described herein has been sponsored by the U.S. Army Combat Capabilities Development Command under cooperative agreement W912CG-19-2-0001. 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**

---

- Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219-243.
- Fahid, F. M., Rowe, J., Kim, Y., Srivastava, S., & Lester, J. (2024, March). Online Reinforcement Learning-Based Pedagogical Planning for Narrative-Centered Learning Environments. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 21, pp. 23191-23199).
- Fahid, F. M., Rowe, J. P., Spain, R. D., Goldberg, B. S., Pokorny, R., & Lester, J. (2021, June). Adaptively scaffolding cognitive engagement with batch constrained deep Q-networks. In *International conference on artificial intelligence in education* (pp. 113-124). Cham: Springer International Publishing.
- Georgila, K., Core, M. G., Nye, B. D., Karumbaiah, S., Auerbach, D., & Ram, M. (2019, May). Using reinforcement learning to optimize the policies of an intelligent tutoring system for interpersonal skills training.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 737-745).

- Johnston, J., Sottolare, R., Sinatra, A. M., & Burke, C. S. (Eds.). (2018). *Building intelligent tutoring systems for teams: What matters*. Emerald Group Publishing.
- Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., & Pineau, J. (2022). Automated data-driven generation of personalized pedagogical interventions in intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 32(2), 323-349.
- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research*, 86(1), 42-78.
- Merrill, M. D. (1983). Component display theory. *Instructional-design theories and models: An overview of their current status*, 1, 282-333.
- Mousavinasab, E., Zarifsanaiy, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1), 142-163.
- Roberts, N., Lenz, T., & Goldberg, B. (2023, July). The GIFT Architectural and Features Update: 2023 Edition. In *Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11)* (p. 3). US Army Combat Capabilities Development Command–Soldier Center.
- Smith, A., Spain, R. D., Rowe, J., Goldberg, B., & Lester, J. (2022, May). Formalizing Adaptive Team Feedback in Synthetic Training Environments with Reinforcement Learning. In *Proceedings of the Tenth Annual GIFT Users Symposium (GIFTSym10)* (p. 117-126). US Army DEVCOM–Soldier Center
- Smith, A., Spain, R., Roberts, N., Goldberg, B., Rowe, J., Mott, B., & Lester, J. (2023, July). Supporting data-driven team feedback and scenario adaptations in GIFT. In *Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11)* (p. 145). US Army Combat Capabilities Development Command–Soldier Center.
- Spain, R., Rowe, J., Goldberg, B., Pokorny, R., Mott, B., & Lester, J. (2021). Automated coaching in synthetic training environments: Developing an adaptive team feedback framework. In *Proceedings of the Ninth Annual GIFT User Symposium (GIFTSym9)*, (pp. 187-199). US Army DEVCOM–Soldier Center.
- Spain, R., Rowe, J., Goldberg, B., Pokorny, R., & Lester, J. (2019). Enhancing learning outcomes through adaptive remediation with GIFT. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (IITSEC)*. Orlando, FL: National Training and Simulation Association.
- Spain, R., Rowe, J., Smith, A., Goldberg, B., Pokorny, R., Mott, B., & Lester, J. (2022). A reinforcement learning approach to adaptive remediation in online training. *The Journal of Defense Modeling and Simulation*, 19(2), 173-193.)
- US Army (2024). The Army Learning Concept for 2030-2040. Retrieved from: <https://adminpubs.tradoc.army.mil/pamphlets/TP525-8-2.pdf>
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4), 197-221.
- Xu, Z., Wijekumar, K., Ramirez, G., Hu, X., & Irey, R. (2019). The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology*, 50(6), 3119-3137.

## ABOUT THE AUTHORS

---

**Andy Smith, Ph.D.** is a Research Scientist in the Center for Educational Informatics at North Carolina State University. He received his M.S and Ph.D. in Computer Science from North Carolina State University, and his B.S. degrees in Computer Science and Electrical and Computer engineering from Duke University. His research is focused on the intersection of artificial intelligence and education, with emphasis on user modeling, game-based learning, and educational data mining.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

**Randall Spain, Ph.D.** is a Research Scientist in the Soldier Effectiveness Directorate at U.S. Army DEVCOM, Soldier Center, Simulation and Training Technology Center (STTC). He holds a Ph.D. in Human Factors Psychology from Old Dominion University and an M.S. in Experimental Psychology. His research focuses on the design and evaluation of advanced training technologies on learning and performance.

**Nicholas Roberts** is a Senior Software Engineer at Dignitas Technologies and the engineering lead for the GIFT project. Nick has been involved in the engineering of GIFT and supported collaboration and research with the intelligent tutoring system (ITS) community for nearly 10 years. Nicholas contributes to the GIFT community by maintaining the GIFT portal ([www.GIFTTutoring.org](http://www.GIFTTutoring.org)) and GIFT Cloud ([cloud.gifttutoring.org](http://cloud.gifttutoring.org)), supporting conferences such as the GIFT Symposium, and technical exchanges with Soldier Center and their contractors. He has also been heavily involved in integrating GIFT into TSS/TMT.

**Jonathan Rowe, Ph.D.** is a Senior Research Scientist in the Center for Educational Informatics at North Carolina State University, Adjunct Assistant Professor in the Department of Computer Science, and Managing Director of the National Science Foundation AI Institute for Engaged Learning. He earned his Ph.D. in Computer Science from North Carolina State University in 2013. His research focuses on designing, developing, and evaluating AI-augmented learning and training technologies, with an emphasis on game-based learning environments, multimodal learning analytics, interactive narrative generation, affective computing, user modeling, and intelligent tutoring systems.

**Bradford Mott, Ph.D.** is a Senior Research Scientist in the Center for Educational Informatics at North Carolina State University. He received his B.S., M.C.S., and Ph.D. in Computer Science from North Carolina State University. Prior to joining North Carolina State University, he worked in the video game industry developing cross-platform middleware solutions used extensively in commercial games and training applications. His research interests include AI and human-computer interaction, with applications in educational technology.

**James Lester, Ph.D.** is a Distinguished University Professor of Computer Science at North Carolina State University, where he is Director of the Center for Educational Informatics. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI). His research on personalized learning technologies ranges from intelligent tutoring systems and game-based learning environments to affective computing, computational models of narrative, and natural language tutorial dialogue.





# **THEME IV: APPLICATIONS OF GIFT**



# A Synthetic Training Environment for Assessing Changes in Team Dynamics with the Generalized Intelligent Framework for Tutoring

Scotty D. Craig, Kevin Gary, Jamie C. Gorman, Vipin Verma, and Robert LiKamWa  
Arizona State University

## INTRODUCTION

---

Most current and future learning and training environments will be dynamic and often distributed across mediums. As work environments increase in complexity and ad hoc coordination is needed, the ability of human-agent teams to coordinate under unplanned circumstances is critical in domains as diverse as defense, healthcare, and education. Meanwhile, team interactions increasingly occur in virtual environments. This learning is a social experience and there are many ways that human learning is influenced through interactions with others. Students can learn socially through collaborative work (Nokes-Malach et al., 2015), teaching other students (Roscoe & Chi, 2007), and observing others (Craig et al., 2009). Team-based components are a good method to facilitate the social aspects of learning. When incorporated within online environments, team components have been shown to improve learning over identical team activities in face-to-face classes (Sohoni et al., 2016). These team-based tasks require teams of individuals with different skill sets to work together to achieve a shared goal. The diverse roles and interdependence are the essence of what the term “team” means (Salas et al., 1992). Teams must conduct tasks together such as planning, sensing, communicating, deciding, and problem solving. In many settings, these tasks require tight coordination and interdependent teamwork. Because of this, there is a need for training in teamwork, monitoring, and feedback to team members (Gilbert et al., 2018; Ouverson et al., 2021). To meet this need, our project will develop experiential team training modules within a synthetic training environment (STE) supported by the Generalized Intelligent Framework for Tutoring (GIFT) to assess and improve the individual team member’s team coordination skills.

## Background

### *Synthetic Training Environments*

While live simulation training can be effective, it is often expensive, time consuming, and dangerous (Levy et al., 2013). Because of this, simulations and virtual systems have a long history within the training area as well as providing effective training (Andrews & Craig, 2015; Swezey & Andrews, 2001). Immersive virtual environments enhance training by allowing users to develop skills in areas such as situational awareness, operational planning, safety protocols (Li et al., 2020), equipment maintenance (Ipsita et al., 2022), site inspections (Eiris et al., 2021), multi-layered data visualization (Bellanca et al., 2019), and tactile equipment handling (Rettinger et al., 2021). These environments offer a rich, experiential platform where individuals or groups can engage deeply with the training material. For instance, they enable personalized training experiences that utilize volumetric 3D data, which can be particularly beneficial in certain situations that demand spatial precision, such as athletic coaching (Eiris et al., 2021). The creation of effective training environments hinges on a seamless integration of user experience design, 3D modeling, and immersive software development, all geared towards achieving specific learning outcomes and skill development.

Within the U.S. Army, these immersive virtual environments have a specific purpose collective training solution that immerses teams of Soldiers in realistic scenarios that target core skills and challenge team

dynamics and are known as Synthetic Training Environments (STEs). STEs are advanced forms of immersive virtual environments that include intelligent tutoring (Craig, 2018; Graesser et al., 2012) and adaptive training services to assist in the preparation and delivery of scenarios that addresses a given teams' or units' training needs (Goldberg et al., 2021). The team process is dynamic and individual just in time feedback can be detrimental to the team learning process (Walton et al., 2014). However, higher levels of role experience and task experience positively impact communication performance (Ouverson et al., 2021) which is essential in team performance. To reliably measure team performance and support effective team training within STEs, metrics of team dynamics are needed (Goldberg et al., 2021).

### ***Team Dynamics***

Team dynamics are critical for supporting adaptive team training and improving team performance. The Generalized Intelligent Framework for Tutoring (GIFT) has previously been used for individual tutoring, and tutoring during team training, and has been fully integrated as a STE. We are currently expanding on GIFT to support team training by integrating team performance sensors that are supported by intelligent tutoring interactions. Previous research on team cognition and its measurement has shown that team training can be monitored by 1) capturing the dynamic nature of team cognition, 2) automating measurement, and 3) conducting measurement in real-time for immediate intervention (Grimm et al., 2023). Coordinated Awareness of Situation by Teams (CAST) is a team SA (situation awareness) measure that evaluates the effectiveness and efficiency of team interaction under "roadblock" scenarios. Team based intelligent tutoring systems must be able to monitor the team environment by receiving feedback on individual and team member performance and use this data to give productive real time feedback.

### **Current System**

Our system would send data from our automated measures of performance to trigger training modules and feedback within GIFT. The team process is dynamic, and individual just-in-time feedback can be detrimental to the team learning process. We discuss our approach for building a generalizable team dynamics measurement framework to assess critical team dynamics competencies including adaptive capacity, resilience, and individual influence. The framework is being used to support experiential team training modules within a STE supported by GIFT. The goal is to assess and improve the ability of team members through objective team measurement that will integrate into the STEEL-R (Synthetic Training Environment Experiential Learning for Readiness) framework (Goldberg et al., 2021) that incorporates virtual and augmented reality to support trainee readiness. Team-based intelligent tutoring systems must be able to monitor the team environment through communication, movement, and often physiological sensors and use this data to give productive real-time feedback (Sottolare et al., 2017).

### **Synthetic Training Environment for Casualty Collection Point (CCP) in Tactical Combat Casualty Care (TC3)**

Our project will develop and leverage a multi-user system that utilizes wirelessly tethered headsets connected to several client computers. These clients interface with a central server responsible for managing the game state and facilitating the coordination of a comprehensive tracking system. We plan to employ a Vicon tracking system, which will track the movements of users' hands, feet, torso, and head, as well as any portable equipment used in the simulation.

The server will also perform inverse kinematics to accurately estimate the full-body skeletons of all users, which is essential for replicating realistic user interactions within the virtual environment. This integration ensures that the physical movements of participants are precisely mirrored in the virtual space, thereby enhancing the overall user experience.

The software development for this system will be carried out using the Unity Engine. Unity is well-suited for this kind of project due to its support for complex simulations and real-time multiplayer interactions, as well as immersive 3D models and animations. This platform provides the necessary tools to build a sophisticated multi-user virtual reality system that aims to study user interaction and immersion.

### **Proposed Team Dynamics Metrics**

As teams encounter perturbations during a scenario, they must reorganize themselves to maintain team effectiveness and exhibit resilience (Gorman et al., 2019). In our project we plan to use learning analytics combined with a layer dynamics approach to measure team effectiveness and use it to provide an adaptive training experience for Casualty Collection Point (CCP) within the Tactical Combat Care (TC3), a complex team-based process that technology is an ideal method for improving training (Milham et al., 2017). During the training simulation we will collect data about verbal communication between teammates along with timestamps, their physical movements in the training space along with gaze vectors, head orientation and speed of motion. We will then use layered dynamics to calculate system entropy as time-series, percent determinism, frequency, average mutual information (AMI), and influence (Grimm et al., 2023). Entropy represents reorganization of system states within and across different layers of systems, trainees, and system controls, with a higher value indicating more variety in system states. Peaks in the entropy over time, especially around the perturbation, are used to evaluate specific team performance measures such as relaxation time, enaction time, adaptation, and resilience. Relaxation time is quantified by the time it takes to achieve extreme values in magnitude of entropy and can be useful in quantifying resilient behavior in teams. Enaction, adaptation, and resilience correspond to the initial, peak, and end times in the resilience curve. Percent Determinism indicates the team member's communication patterns. A value of 100 would mean a completely predictable communication pattern while zero would mean complete randomness. AMI measures the influence of one team member's activity or communication in reducing the uncertainty of the entire team's activity or communication patterns. A highly influential team member such as a leader would have a higher AMI compared to others. These team measures will then be utilized for formative as well as summative assessments. As part of formative assessment, they will be used to adapt the simulation to improve the team dynamics, provide feedback, and provide required remediation during the training scenario.

### **Preliminary System Architecture**

We will work on creating a system architecture that will be required for the reliable operation of the software tool needed for this project. We will support bi-directional systems-of-systems integration between the virtual reality-enhanced learning environment, the GIFT platform (GIFT Developer Guide, 2023; Hoffman et al., 2021), and the overall STEEL-R system (Hernandez et al., 2022). Our envisioned training environment can generate rich data from events and sensors within the environment. Fortunately, GIFT provides clean integration interfaces for this kind of data. We are undertaking a scenario-driven analysis to flesh out the desired granularity of the information that should be sent downstream. Competency-based assessments and other specific forms of measurement data will flow from our environment into the STEEL-R system. GIFT and STEEL-R provide multiple patterns of integration (internal/external) that could be followed, and we are evaluating exemplars and best practices for this integration. Specifically, we plan to follow the integration pathway pattern established for STEEL-R by sending xAPI statements downstream over the LRS (learner record store) Pipeline.

We are also working to determine what information and event flow might be bi-directional. For example, how will learner profiles be exposed to the virtual learning environment to ensure this environment has sufficient (and only necessary) learner-specific information? What events might GIFT/STEEL-R need to communicate to the learning environment as it is running? This architecture integration complexity is

compounded by the team-oriented focus of our training scenarios, requiring multiple coordinating event streams to pass data and control information in both directions.

### Integrating with the GIFT Platform

The predominant (but not a hard requirement) integration pattern between GIFT and new training applications is to create an “Interop Plugin” (IP) between GIFT’s Gateway module and the training application. System integrators write both sides of the integration and utilize a “drop-in” plugin style to deploy these adapters on both sides (see Figure 1). Well-defined Java interfaces and abstract class types (e.g. *AbstractInteropInterface*) are provided on the GIFT side with appropriate configuration files in order to extend GIFT to identify a new Interop Plugin at startup (GIFT Developer Guide, 2023). We will refer to this side as the GIFT-IP.

On the training application side, a complimentary component to the Interop Plugin must be deployed. We will refer to this side as the TA-IP. As GIFT is agnostic to the underlying technology supporting a training application, the manner in which this side is constructed is up to the training app developer. The underlying communication protocols used between the GIFT-IP and the TA-IP are also up to the training application developer since they write both sides of the integration. Further, more than one communication channel may be created between the GIFT-IP and the TA-IP to accommodate data and control messaging requirements. How we intend to utilize this is discussed in the next section.

Once communication is established with GIFT, a pathway exists to utilize the various modules and services provided by GIFT, as well as access the larger STEEL-R backplane.

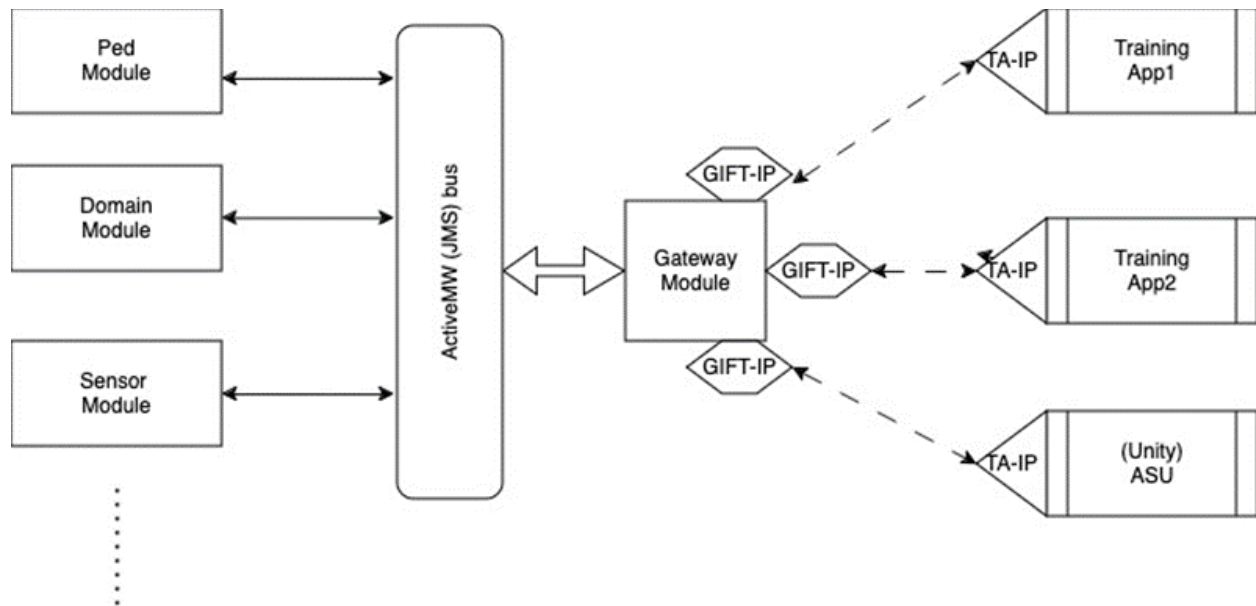


Figure 1. Envisioned high-level architecture

### Communication with the GIFT Platform

The GIFT platform provides tremendous flexibility for training application developers and system integrators to select communication protocols for data and control message exchange. Examples include gRPC, XML-RPC, GWT-RPC, binary low-level sockets and more. The semantics of the payloads in these hybrids are hybrid, meaning GIFT does define some lifecycle control message types, and interop plugins may define their own. For our purposes, we foresee handling the message types shown in Table 1.

**Table 1. Envisioned abstract message types**

Message	Type	Direction	Description
Learner Events	data	TA-IP to GIFT-IP	Our training application will filter relevant training apps within the app, and push these events to GIFT.
Perturbations	control	GIFT-IP to TA-IP	The application will introduce perturbations (state changes on individual objects in the scene) at various points during the simulation.
Lifecycle	control	GIFT-IP to TA-IP	GIFT defines a set of standard lifecycle event types (SIMAN messages) which may be augmented by a training app. Given the team-oriented nature of our simulation exercise these will be revisited later.
Remediations	control	GIFT-IP à TA-IP	Learner-focused directives issued as part of the adaptive control cycle based on learner competency measures evaluated in quasi-real time.

At present we expect to utilize two communication channels, one for Learner Events (data) which will be pushed at a frequency that will support roundtrip evaluation through the competency measures module (see below). This channel will be dedicated and optimized as we expect relatively “chatty” traffic with specialized quality of service measures (i.e. speed, reliability). The second channel will handle the control-type messages that flow from GIFT to the TA. We do not foresee needing more than one channel, though we will experiment with our architectural prototype to ensure performance requirements are met.

Fortunately, on the TA-IP side, GIFT already provides an SDK that includes templates and examples for communicating within Unity to a GIFT server. We intend to extend this starting point in our platform.

**Integrating with the GIFT/STEEL-R backplane**

### ***Level 1 Backplane Integration***

GIFT's Gateway module is the entry point for training application integration, and eventually to the entire STEEL-R backplane (Goldberg et al., 2021). Our initial "Level 1" integration is focused on the GIFT-IP to TA-IP integration described above, but messages do need to route through the backplane. Initially we intend to show that the information produced by our training application can be transformed to normalized messages that go on the GIFT JMS (ActiveMQ) message bus. Further, we will initially develop our learning competency measures analytical tool to sit on the message bus (unless Learner Events prove too chatty for this shared bus). This tool will compute our quasi-real time competency measures and recommend remediations back into the learning environment in the training application.

### ***Future Integration Levels***

Eventually the initial tool we create will need to be more tightly integrated with the STEEL-R competency framework, while maintaining the unique aspects of team-based learning and adaptation in the training application environment. We intend to tackle this "Level 2" integration once the Level 1 is completed. The Adaptive Learning Service API added in GIFT release 2021 (GIFT Developer Guide, 2023) is a possible starting point for supporting the adaptive microlearning cycle of our training application.

We are also presently working on decoupling GIFT deployment modules so they support a containerized approach. We intend to develop our multiple competencies as individual microservices that can evolve independently as we gain a better understanding of their effectiveness in the training environment.

## **CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH**

---

The project will culminate with an integrated system using GIFT that will train TC3. This system will allow for small teams to experience TC3 scenarios under multiple perturbations within the environment which will allow for our team dynamic algorithms to detect team performance and cohesion. These scenarios will be implemented within an immersive virtual environment created to implement a STE for TC3. Each session will introduce perturbations that interfere with normal team interactions. Perturbations are operationalized as novel situations that require adaptive team communication and coordination. Team competencies assessed using perturbation training have been found to be valid indicators of team success when transferring training to novel contexts. The scenario play through will provide data for an interactive after action review session using GIFT's interactive tutoring functionality to provide training on team performance.

## **ACKNOWLEDGEMENTS**

---

The research described herein has been sponsored by the U.S. Army Combat Capabilities Development Command under cooperative agreement W912CG-23-2-0002. 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**

---

Andrews, D. H., & Craig, S. D. (Eds.). (2015). Readings in training and simulation (Vol. 2): Research articles from 2000 to 2014. Santa Monica, CA: Human Factors and Ergonomics Society.



## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Bellanca, J. L., Orr, T. J., Helfrich, W. J., Macdonald, B., Navoyski, J., & Demich, B. (2019). Developing a virtual reality environment for mining research. *Mining, metallurgy & exploration*, 36, 597-606.
- Craig, S. D. (Ed.). (2018). *Tutoring and intelligent tutoring systems*. New York, NY, USA: Nova Science Publishers.
- Craig, S. D., Chi, M. T. H., & VanLehn, K. (2009). Improving classroom learning by collaboratively observing human tutoring videos while problem solving. *Journal of Educational Psychology*, 101, 779-789.
- Eiris, R., Wen, J., & Gheisari, M. (2021). iVisit-Practicing problem-solving in 360-degree panoramic site visits led by virtual humans. *Automation in Construction*, 128, 103754.
- GIFT Developer's Guide (2023). GIFT Developer's Guide 2023-1. [https://www.gifttutoring.org/projects/gift/wiki/Developer\\_Guide\\_2023-1](https://www.gifttutoring.org/projects/gift/wiki/Developer_Guide_2023-1)
- Gilbert, S. B., Slavina, A., Dorneich, M. C., Sinatra, A. M., Bonner, D., Johnston, J., ... & Winer, E. (2018). Creating a team tutor using GIFT. *International Journal of Artificial Intelligence in Education*, 28, 286-313.
- Goldberg, B., Owens, K., Hellman, K., Robson, R., Blake-Plock, S., Hoffman, M. & Gupton, K. (2021). Forging Competency and Proficiency through the Synthetic Training Environment with an Experiential Learning for Readiness Strategy. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. Orlando, FL USA. <https://gifttutoring.org/attachments/download/4295/21332.pdf>.
- Gorman, J. C., Demir, M., Cooke, N. J., & Grimm, D. A. (2019). Evaluating sociotechnical dynamics in a simulated remotely-piloted aircraft system: A layered dynamics approach. *Ergonomics*, 62(5), 629-643.
- Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA educational psychology handbook*, Vol. 3. Application to learning and teaching (pp. 451-473). American Psychological Association. <https://doi.org/10.1037/13275-018>
- Grimm, D. A., Gorman, J. C., Cooke, N. J., Demir, M., & McNeese, N. J. (2023). Dynamical Measurement of Team Resilience. *Journal of Cognitive Engineering and Decision Making*, 17(4), 351-382.
- Hernandez, M., Blake-Plock, S., Owens, K., Goldberg, B., Robson, R., Center, S., & Ray, F. (2022). Enhancing the total learning architecture for experiential learning. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.
- Hoffman, M., Goldberg, B., & Brawner, K. (2021, May). The GIFT Architecture and Features Update: 2021 Edition. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym10)* (p. 11).
- Ipsita, A., Erickson, L., Dong, Y., Huang, J., Bushinski, A. K., Saradhi, S., ... & Ramani, K. (2022, April). Towards modeling of virtual reality welding simulators to promote accessible and scalable training. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1-21).
- Levy, M., Koch, R. W., & Royne, M. B. (2013). Self-reported training needs of emergency responders in disasters requiring military interface. *Journal of Emergency Management*, 11(2), 143-150.
- Li, M., Sun, Z., Jiang, Z., Tan, Z., & Chen, J. (2020). A virtual reality platform for safety training in coal mines with AI and cloud computing. *Discrete Dynamics in Nature and Society*, 2020, 1-7.
- Milham, L., Phillips, H., Ross, W., Townsend, L., Riddle, D., Smith, K., Butler, P., Wolf, R., Irizarry, D., Hackett, M., & Johnston, J. (2017). Squad-level training for tactical combat casualty care: Instructional approach and technology assessment. *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 14(4), 345-360. <https://doi.org/10.1177/1548512916649075>
- Ouverson, K. M., Ostrander, A. G., Walton, J., Kohl, A., Gilbert, S. B., Dorneich, M. C., ... & Sinatra, A. M. (2021). Analysis of communication, team situational awareness, and feedback in a three-person intelligent team tutoring system. *Frontiers in Psychology*, 12, 553015. <https://doi.org/10.3389/fpsyg.2021.553015>
- Nokes-Malach, T. J., Richey, J. E., & Gadgil, S. (2015). When is it better to learn together? Insights from research on collaborative learning. *Educational Psychology Review*, 27(4), 645-656.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Rettinger, M., Müller, N., Holzmann-Littig, C., Wijnen-Meijer, M., Rigoll, G., & Schmaderer, C. (2021, May). Vr-based equipment training for health professionals. In Extended abstracts of the 2021 CHI conference on human factors in computing systems (pp. 1-6).
- Roscoe, R. D., & Chi, M. T. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, 77(4), 534-574.
- Salas, E., Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an understanding of team performance and training. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 3-29). Westport, CT, US: Ablex Publishing.
- Sohoni, S., Mar, C., & Craig, S. D. (2016). Comparing cooperative learning in online and in-person versions of a microprocessors course. In L. Gossage (Ed.), *Proceedings of the 2016 American Society for Engineering Education Pacific Southwest Section Conference* (pp. 257-268). American Society for Engineering Education.
- Sottolare, 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*, 1-19.
- Swezey, R. W., & Andrews, D. H. (2001). *Readings in training and simulation: A 30-year perspective*. Santa Monica, CA: Human Factors and Ergonomics Society.
- VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., et al., (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, 15(3), 147–204.
- Walton, J., Dorneich, M. C., Gilbert, S., Bonner, D., Winer, E., & Ray, C. (2014). Modality and timing of team feedback: Implications for GIFT. In *Proceedings of the Second Annual GIFT User Symposium* (pp. 190-198).

## ABOUT THE AUTHORS

---

**Scotty Craig, Ph.D.** is an Associate Professor in Human Systems Engineering with the Ira A. Fulton Schools of Engineering at Arizona State University. Dr. Craig is the Director of Research and Evaluation for the ASU Learning Engineering Institute and Director of the ASU Advanced Distributed Learning Partnership Laboratory. His work is at the intersection of cognitive science, user science and science of learning working to develop and evaluate learning ecosystems.

**Kevin Gary, Ph.D.** is an Associate Professor in Software Engineering in the School of Augmented Intelligence with the Ira A. Fulton Schools of Engineering at Arizona State University. Dr. Gary's interests are in Agile and Lean Software Engineering, Software Engineering for Education and Healthcare, and Software Engineering Education.

**Jamie Gorman, Ph.D.** is a Professor in Human Systems Engineering with Ira A. Fulton Schools of Engineering at Arizona State University. Dr. Gorman is an expert in modeling and measuring coordination dynamics in human and human-machine teams. His thriving research portfolio includes psychology topics, investigating human and artificial intelligence dynamics and human-machine teaming for space-based missions.

**Vipin Verma, Ph.D.** is an Assistant Research Scientist at Decision Theater (DT) in the Knowledge Enterprise, Arizona State University. His research interests include Educational Games, Simulation, Gamification, Assessment, Psychometrics, and Affective Computing.

**Robert LiKamWa, Ph.D.** is an Associate Professor at Arizona State University, appointed in the School of Arts, Media and Engineering (AME) and the School of Electrical, Computer and Energy Engineering (ECEE). LiKamWa directs Meteor Studio, which explores the research and design of software and hardware for mobile Augmented Reality, Virtual Reality, Mixed Reality, and visual computing systems, and their ability to help people tell their stories.

# Enhancing Data Science Courses Pedagogy through GIFT-Enabled Adaptive Learning Pathways

Fadjimata I. Anaroua<sup>1</sup>, Qing Li<sup>2</sup>, and Hong Liu<sup>1</sup>  
 Embry-Riddle Aeronautical University<sup>1</sup>, University of North Texas<sup>2</sup>

## INTRODUCTION

Over the past decade, the educational landscape has experienced a surge of online learning and instructional platforms (Liu et al., 2020). This remarkable surge can be attributed to a confluence of factors, including the rising demand for higher education opportunities, the shortage of available teaching staff, and the rapid advancements in information technology and artificial intelligence (AI) capabilities. AI remained a niche area of research with limited practical applications in education for over half a century (Bhutoria, 2022; Chen et al., 2020; Roll & Wylie, 2016) from 1950 to 2010. However, in recent years, the advent of Big Data and advancements in computing power have propelled AI into the educational mainstream (Alam, 2021; Chen, Chen, & Lin, 2020; Chen, Xie et al., 2020; Hwang et al., 2020). Today, the rise of machine learning, deep learning, automation, together with advances in big data analysis has sparked novel perspectives and explorations around the potential of enhancing personalized learning, a long-term educational vision of technology-enhanced course options to meet student needs (Grant & Basye, 2014).

Fostering personalized learning necessitates the development of digital learning environments that dynamically adapt to individual learners' knowledge, prior experiences, and interests, while effectively and efficiently guiding them towards achieving desired learning outcomes (Spector, 2014, 2016). AI-powered technologies have made it possible to analyze data generated by learners and provide instruction that matches their learning performance. Through learning analytics and data mining techniques, large datasets collected are analyzed and processed to uncover learners' unique learning characteristics, often referred to as learner profiling (Tzouveli et al., 2008). Subsequently, leveraging AI algorithms, the learning content is tailored, and personalized learning paths are designed to align with each learner's identified needs and preferences, thereby facilitating personalized learning experiences. See Figure 1 for the advanced digital learning environment.

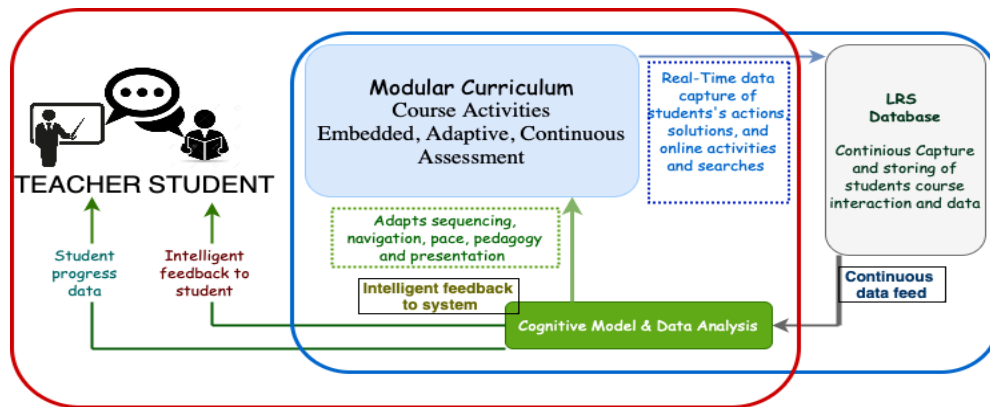


Figure 1. Advanced Digital Learning Environment

The data-driven personalization allows educators to create targeted interventions, address learning gaps, and provide enriched learning experiences for all students, moving from static learning environments to

dynamic, adaptive systems that respond in real-time to learner data (Imhof et al., 2020). However, this AI-driven pedagogical shift requires profound transformations in conventional educational methodologies. Educators must transition from a one-size-fits-all instructional model to a learner-centric paradigm, where teaching strategies, content delivery, and assessment methods are dynamically tailored to each individual's unique needs and learning trajectories. Such a shift requires a deeper understanding of learner characteristics, cognitive processes, and the effective integration of AI-driven analytics and adaptive systems into the classroom experience (Walkington & Bernacki, 2020). Moreover, the role of educators evolves from being mere content providers to facilitators and guides, leveraging AI-driven insights to create personalized learning pathways, provide targeted feedback, and foster self-directed learning (Kim et al., 2014; Kompen et al., 2019). Collaborative learning experiences, where AI systems augment and support human interactions, become increasingly important, enabling learners to co-construct knowledge and develop critical thinking and problem-solving skills.

Assessment practices must also adapt, moving beyond traditional summative assessments to incorporate continuous, formative, and adaptive assessments that provide real-time feedback and inform the personalization of learning experiences. The integration of AI-driven learning analytics and predictive modeling can help identify learners' strengths, weaknesses, and potential learning obstacles, enabling proactive interventions and targeted support.

To address these limitations, the Generalized Intelligent Framework for Tutoring (GIFT) was developed, adopting a modular, service-oriented architecture to promote modularity, reusability, and domain independence (Sottolare & Brawner, 2018). Originally designed for the U.S. Army to understand and adapt to the needs of individual learners and to guide instruction in real time to optimize learning (Sottolare et al., 2017), GIFT aimed to improve scalability, enhance adaptivity through advanced learner modeling and instructional strategies, and enable interoperability with emerging standards like xAPI.

By providing a generalized and extensible framework, adaptive learning environments enabled by GIFT have facilitated self-regulated learning in an unobtrusive yet personalized manner, tailored to the unique needs of individual learners and small teams, enabling the necessary research and development leading to standardized, adaptive tutoring systems that are accessible, flexible, affordable, and easy to develop and utilize (Sottolare et al., 2023).

## LITERATURE REVIEW

---

Adaptive learning technologies are grounded in the principle that education should be as unique as the individual learner. These systems dynamically adjust the educational content and instructional strategies based on real-time assessments of learner performance and behavior. This adaptivity is made possible through a combination of data analytics, AI, and sophisticated algorithms. The goal is to create a personalized learning experience that aligns with each learner's cognitive abilities, preferences, and knowledge levels, catering to a wide spectrum of learning styles. In the ever-evolving realm of education, the advent of adaptive learning technologies marks a pivotal shift, heralding a new era in the delivery of educational content and instructional strategies (Martin et al., 2020).

Uniting the principles of personalized and adaptive learning, personalized adaptive learning emerges as a technology-empowered effective pedagogy that can adaptively adjust teaching strategies based on real-time monitoring of learners' differences and changes in the core elements of individual characteristics, individual performance, personal development, and adaptive adjustment (Peng et al., 2019). Adaptive learning technology provides personalized learning at scale by assessing learners' current skills and knowledge, providing feedback and content, and then constantly monitoring progress, utilizing learning algorithms that provide real-time updates and the necessary tools to improve student learning (Taylor et al., 2021).

The rise of personalized adaptive learning approaches is shifting the educational paradigm from a traditional instructor-centric model to a student-centric one. According to Dockterman's (2018) overview, learners tend to absorb material more effectively when the instruction is tailored to their individual requirements, and the pedagogy of personalization acknowledges that every student is unique. However, putting this learner-focused teaching methodology into practice necessitates the use of technological solutions and platforms capable of dynamically adapting to the distinct needs, capabilities, and learning preferences of each individual student.

Researchers have proposed frameworks and models for designing personalized learning paths and adaptive instructional strategies based on learner profiles, competency progression, and flexible learning environments (Peng et al., 2019). Efforts have also been made to design adaptive cloud-based educational systems to integrate adaptive learning with augmented reality (Marienko et al., 2020).

However, most existing studies on personalized adaptive learning instruction have focused on K-12 classrooms, with limited exploration of the reform effort in tertiary-level personalized adaptive learning (Christodoulou & Angeli, 2022; Taylor et al., 2021; Ryoo & Winelmann, 2021). Besides, there is a lack of targeted investigations into specific disciplines, which may have distinct pedagogical considerations. There is a notable research gap in exploring personalized adaptive learning environments specifically tailored to higher education contexts and disciplines like data science courses. This gap presents an opportunity for further research to address the unique needs and challenges of implementing personalized adaptive learning in tertiary-level data science education.

As a modular, open-source architecture that is designed to facilitate the development and deployment of adaptive instructional systems and intelligent tutoring systems (Sottolare et al., 2013), GIFT itself does not generate adaptive learning pathways, but provides a framework for integrating various components such as sensor, trainee, pedagogical, LMS (learning management system; Hoffman & Ragusa, 2015), to enable personalized and adaptive learning experiences. Specialized authoring tools like Articulate can be used to create interactive and multimedia-rich learning content, built in three major tools – Presenter, Engage, and Quizmaker (Martin & Martin, 2015), which can then be integrated into GIFT-enabled adaptive learning systems. The content created in Articulate could be mapped to specific learning objectives, competencies, and instructional strategies within GIFT's domain model. This would allow GIFT to dynamically select and present relevant content based on the learner's performance, preferences, and needs.

The Moodle LMS, widely used in online teaching and learning in STEM education (Gamage et al., 2022), can also be integrated with GIFT to facilitate the delivery and tracking of adaptive learning experiences (Despotović-Zrakić et al., 2012). GIFT could potentially interface with Moodle to retrieve learner data, such as performance metrics and progress, and use this information to generate personalized learning paths and adaptive instructional strategies. Conversely, GIFT could push adaptive content and assessments to Moodle for learners to access and complete.

While the framework of adaptive learning systems is helpful, it remains elusive how we can practically support the necessary multi-level and interdependent nature among systems to streamline their integration and tailor them to learners' needs at different levels and in different contexts. To address this, we explore the integration of GIFT with learning and authoring tools, aiming to create a robust adaptive learning ecosystem for data science education. By harnessing the strengths of GIFT's adaptive learning engine combined with Articulate and Moodle's extensive repository of educational resources, we aim to deliver personalized learning pathways that are not only aligned with individual learning preferences and needs but also scalable across diverse educational settings. This paper introduces a novel adaptive learning model for data science education and presents analytical frameworks that enhance teaching practices. We demonstrate the integration of GIFT with Moodle and Articulate, showcasing how adaptive learning technologies can provide tailored and efficient learning experiences. Built on the GIFT platform, our findings highlight the

transformative impact of these technologies in data science education and offer insights applicable to other fields. This research contributes to discussions on incorporating adaptive learning into existing educational frameworks, paving the way for personalized and effective learning experiences in the digital age.

## INITIAL DESIGN

---

The initial design for investigating the efficacy and impact of the integrated adaptive learning system, comprising Articulate, Amazon Simple Storage Service (S3), GIFT and Moodle, adopts a mixed-methods approach to provide both quantitative and qualitative insights into its effectiveness within data science education. The quantitative component involves a controlled experiment where participants are divided into two groups: one using the traditional learning model and the other utilizing the new adaptive learning ecosystem. Key performance indicators such as engagement levels, learning outcomes, and time to competency, will be rigorously measured and analyzed. Concurrently, the qualitative aspect will employ surveys, interviews, and focus groups to gather in-depth feedback from students and educators regarding their experiences, perceived benefits, and any challenges encountered. This comprehensive approach aims to not only quantify the benefits of the adaptive learning system but also to understand the nuanced experiences of its users, thereby offering a holistic view of its potential to transform data science education.

## INTEGRATION TOOLS AND PROCESSES

---

The main integration process of the proposed adaptive learning ecosystem, encompassing Articulate for content creation, Amazon S3 for content hosting, and GIFT and Moodle for content delivery, is a meticulously designed workflow that ensures seamless interoperability and maximizes the effectiveness of adaptive learning in data science education.

**Content Creation with Articulate:** The process begins with the development of interactive and adaptive learning materials using Articulate, a sophisticated authoring tool known for its ability to create engaging multimedia content. Educators and content creators design courses with adaptive pathways, incorporating various learning activities, assessments, and multimedia elements tailored to diverse learning styles and needs. In the context of e-learning, the concept of "branched" is frequently mentioned. This typically refers to "branched scenarios" or "branched e-learning," which describes a course structure that allows learners to navigate or be redirected through various pathways, also known as "branches". Courses with branching, or non-linear designs, offer a more tailored experience for learners. By integrating branching scenarios, learners can navigate through simulations of real-world scenarios, discovering how their decisions can result in diverse outcomes. Alternatively, a course design offering varied pathways to accommodate individuals/teams with distinct roles or varying degrees of background knowledge is also possible (Legault, n.d.).

We utilized Articulate Storyline 360's robust features, specifically variables and triggers, to develop customized branching scenarios that respond to a learner's progress within the e-learning course. The method employed involves leveraging the platform's quiz and results slides to establish scoring variables. Subsequently, we implemented conditional triggers to navigate users to different course branches based on their accumulated points, effectively modifying variable values to create a tailored learning journey. See Figure 2 for Storyline 360's variables and triggers for the adaptive course concept.

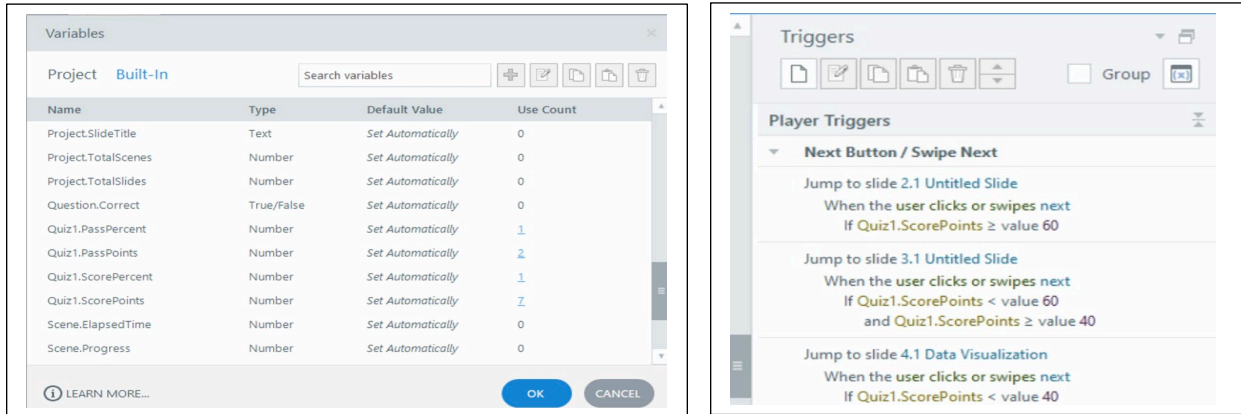


Figure 2. Storyline 360's Variables and Triggers for the Adaptive Course Concept

**Hosting on Amazon S3:** Once the content is developed, it is uploaded to Amazon S3, a reliable and scalable cloud storage service and managed through CloudBerry Explorer. This step involves configuring storage buckets in S3, ensuring the content is securely stored and readily accessible. Amazon S3 serves as a centralized repository for the adaptive learning materials, offering robust data protection, high availability, and seamless content delivery. According to Slim (2021), CloudBerry Explorer is used as an intuitive file explorer that helps manage an Amazon S3 account just as if it was one more folder on the local drive. The program features a double-pane interface and works like an FTP client, with each window devoted to a location. CloudBerry Explorer supports multiple S3 accounts and lets users open multiple connections, which can then easily be managed through interface tabs. A web URL for the hosted content on Amazon S3 for our initial design is generated from the CloudBerry interface.

**Integration within GIFT and Moodle Platform:** The final step involves integrating the hosted content with GIFT, as our advanced adaptive learning engine. This integration is facilitated through APIs and specific integration protocols that enable GIFT to fetch and present the adaptive content stored in Amazon S3 to learners. We have deployed our GIFT software on an Amazon AWS EC2 instance and integrated our adaptive content. The web URL generated from Amazon S3 will be embedded within the GIFT platform using the "Web address" course object. The subsequent step involves connecting our hosted GIFT as an "External Tool" to the Moodle LMS platform using Learning Tools Interoperability (LTI). This integration will enable students and teams to access the adaptive content, facilitating the initial design and assessment of teamwork competence. GIFT dynamically adjusts the learning content based on real-time analysis of learner interactions, performance data, and feedback, providing a personalized learning experience for each student or team. Moreover, the Learning Record Store (LRS) will capture the xAPI statements of interactions from our GIFT-Moodle integration, ensuring detailed tracking and recording of all learning activities.

## CONCEPT IMPLEMENTATION AND TESTING

The implementation of the adaptive learning ecosystem, integrating Articulate for content creation, Amazon S3 for hosting, GIFT for adaptive learning delivery, and Moodle for additional resources and activities, encompasses several successful key phases. Each phase focuses on a specific aspect of the system's

deployment, from the technical setup and configuration to the creation of adaptive learning pathways and the enhancement of user experience.

Through Articulate's intuitive interface, users can design intricate branching scenarios that present learners with decision points, leading to different paths based on their responses. This capability enhances the interactivity and personalization of courses, making learning more engaging and closely aligned with each learner's needs. See Figure 3 for the initial adaptive path branching course content with triggers and variables.

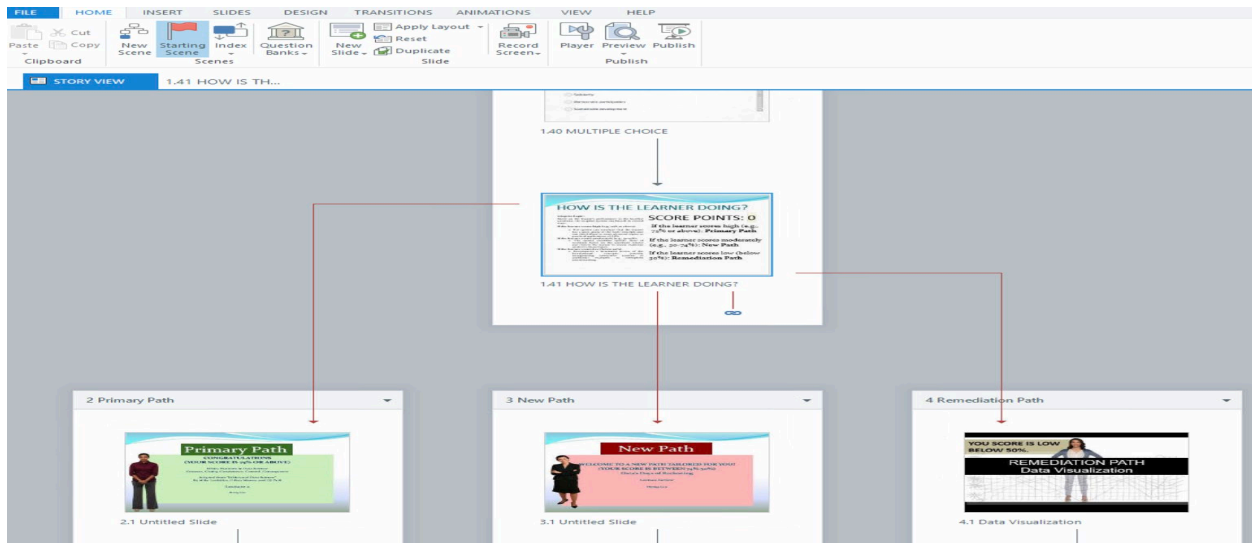


Figure 3. Initial Adaptive Path Branching Course Content with Triggers and Variables

Moodle's extensive repository of learning resources and interactive activities is integrated to enrich the ecosystem. Key steps include content synchronization to ensure that GIFT resources are accessible within the Moodle platform, providing a unified learning experience. The activity tracking is for Integrating Moodle's activity data into the adaptive learning analytics to be captured by the Learning Record Store (LRS), allowing GIFT to adjust learning pathways based on interactions within Moodle. Overall, the implementation of this adaptive learning ecosystem is a multi-faceted process that requires careful planning, coordination among various technology platforms, and a steadfast focus on the learner's experience. The ultimate goal is to create a dynamic, engaging, and personalized learning environment that effectively supports the diverse needs of data science students and teams. See Figure 4 for an example of the adaptive content integration with GIFT.



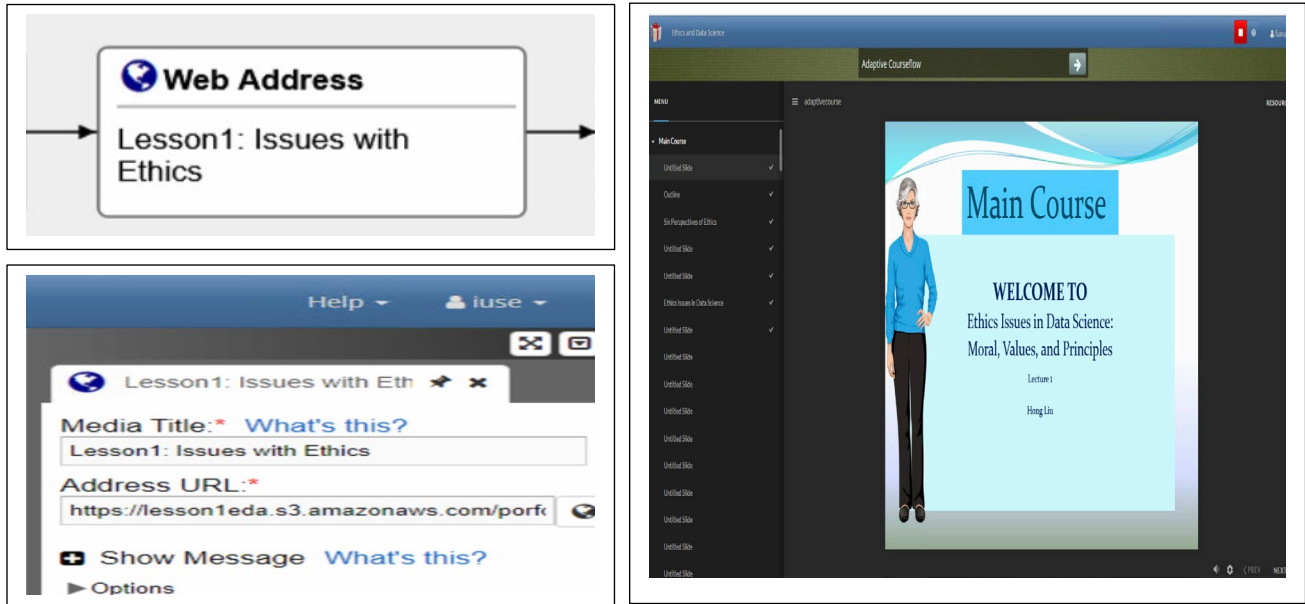


Figure 4. Adaptive Content Integration to GIFT

## EVALUATION AND ANALYSIS

### Evaluation Framework and Data Collection Methods

An integral component of our adaptive learning ecosystem is the LRS, a specialized database designed for storing, retrieving, and analyzing learning data in compliance with the Experience API (xAPI) standard. The LRS serves as the backbone for data-driven decision-making within the system, enabling a nuanced understanding of learner interactions, behaviors, and outcomes. This section delineates the strategic incorporation of the LRS into our ecosystem, detailing its role, configuration, and the benefits it brings to adaptive learning in data science education.

Setting up the LRS involves selecting a robust and scalable LRS platform that can support the anticipated volume of learning data and integrate seamlessly with our existing infrastructure. The chosen LRS (Veracity Learning), is configured to communicate with GIFT, Moodle, and the content hosted on Amazon S3, ensuring that learning activities are accurately captured and logged in real-time.

The LRS plays a pivotal role in shaping adaptive learning pathways within GIFT. By analyzing the comprehensive dataset stored within the LRS, the identification of patterns, trends, and learning gaps at both individual and cohort levels can be possible through GIFT. This data-driven insight allows GIFT content creators to dynamically adjust learning pathways, content recommendations, and instructional strategies, ensuring that each learner receives a personalized and optimally challenging learning experience. Figure 5 and Figure 6 illustrate the overview of a class of learners' learning activities and an individual learner's learning activities, respectively.

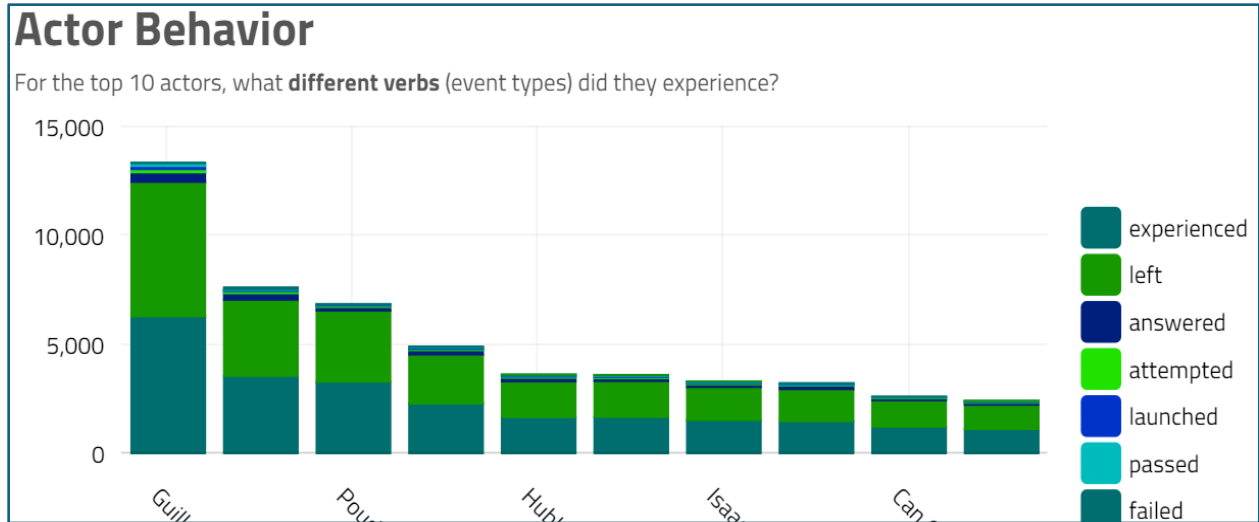


Figure 5. LRS overview of the learners' learning activities sorted from the most active to least active learners.

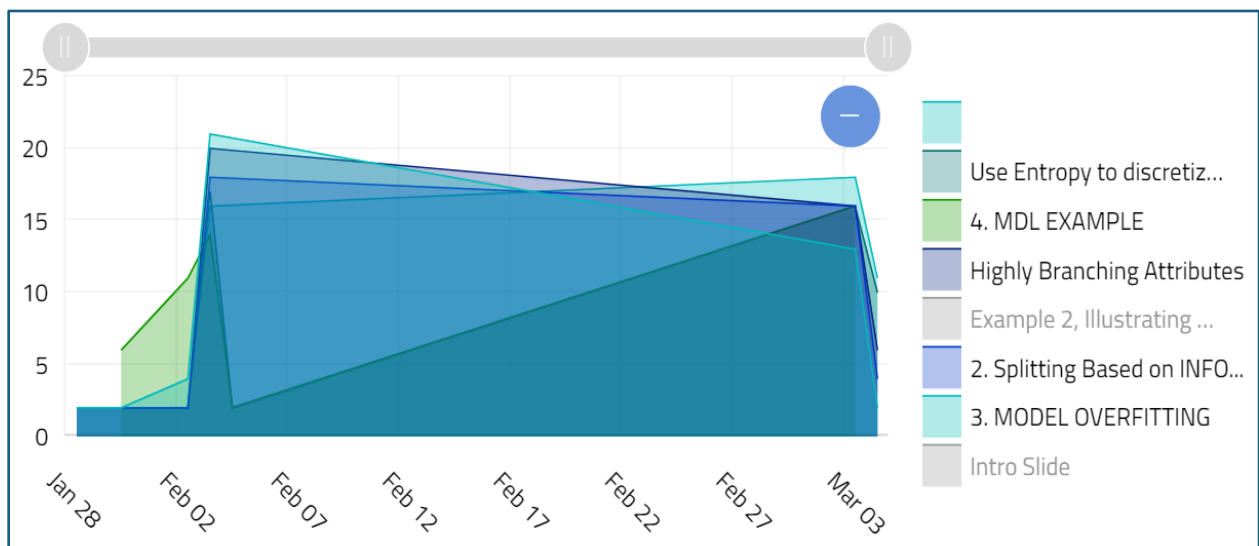


Figure 6. LRS overview of an individual learner's learning activities and content details.

Moodle's vast repository of educational resources and interactive activities is leveraged by feeding data on learner interactions within Moodle into the LRS. This integration ensures that learners' experiences with Moodle's resources contribute to the adaptive learning pathways in GIFT, creating a cohesive and comprehensive learning experience that spans multiple platforms.

The LRS functions as the central repository for learning records generated by learners as they interact with the adaptive learning content hosted on Amazon S3 and delivered through GIFT. These records encapsulate a wide array of data points, including learner responses, time spent on activities, assessment scores, and navigational paths, among others. By adhering to the xAPI specifications, the LRS ensures that learning data from various sources, including Articulate-authored content and Moodle activities, are uniformly structured and semantically rich, facilitating advanced analytics and insights.

## CONCLUSIONS AND FUTURE WORK

---

This paper provides a comprehensive framework for integrating an adaptive learning system utilizing Articulate, Amazon S3, GIFT, and Moodle to enhance data science education. The mixed-methods approach used in the study effectively combines quantitative and qualitative data to assess the system's impact. Quantitatively, the use of controlled experiments helped measure engagement levels, learning outcomes, and time to competency. Qualitatively, surveys, interviews, and focus groups offered valuable insights into the users' experiences, benefits, and challenges encountered. This holistic approach not only quantifies the benefits but also deeply understands the nuanced experiences of learners, making a strong case for the potential transformative impact of this integrated adaptive learning system on data science education.

To date, two adaptive courses, Data Visualization and Data Mining were developed and delivered to the 56 students at Embry-Riddle Aeronautical University. Future research could expand in several directions such as a broader implementation for scaling the system to include a wider variety of courses and academic disciplines beyond data science to examine its adaptability and effectiveness across different educational contexts. Longitudinal Studies can also be conducted for long-term studies to assess the sustained impact of adaptive learning systems on students' academic performance and retention rates over time. Advanced Analytics is another focus where utilizing more sophisticated data analytics techniques will help to delve deeper into the data collected through the LRS, which could help refine the adaptive learning models further and AI enhancements. Incorporating more advanced AI technologies, such as machine learning and natural language processing may help enhance the personalization capabilities of the learning system.

Finally, the Team Competence Assessment will be further explored for developing and integrating metrics and tools within the adaptive learning system to assess and enhance team competence systematically. This involves creating frameworks that not only measure individual learning outcomes but also evaluate collaborative skills, communication, and problem-solving abilities in a group setting. Future studies could explore the dynamics of team-based learning environments and devise methods to optimize team interactions and performance in real-time. This could also include the use of AI to analyze team interactions and provide feedback to improve group cohesion and effectiveness.

## ACKNOWLEDGEMENTS

---

This research was sponsored by the National Science Foundation (NSF) of the USA under the IUSE Grant NSF Grant 2142514 to Embry-Riddle Aeronautical University and NSF Grant 2142327 to University of North Texas, 2022-2025. The authors would like to thank Veracity Learning Inc. for donating the LRS software system and the colleagues in GIFT research and development for their free online technical support.

## REFERENCES

---

- Alam, A. (2021). Possibilities and apprehensions in the landscape of artificial intelligence in education. In *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)* (pp. 1-8). IEEE.
- Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, 3, 100068.
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, 8, 75264-75278.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, 100002.
- Christodoulou, A., & Angeli, C. (2022). Adaptive Learning Techniques for a Personalized Educational Software in Developing Teachers' Technological Pedagogical Content Knowledge. In *Frontiers in Education* (Vol. 7, p. 789397). Frontiers.
- Despotović-Zrakić, M., Marković, A., Bogdanović, Z., Barać, D., & Krčo, S. (2012). Providing adaptivity in Moodle LMS courses. *Journal of Educational Technology & Society*, 15(1), 326-338.
- Dockterman, D. (2018). Insights from 200+ years of personalized learning. *npj Science of Learning*, 3(1), 15.
- Gamage, S. H., Ayres, J. R., & Behrend, M. B. (2022). A systematic review on trends in using Moodle for teaching and learning. *International journal of STEM education*, 9(1), 9.
- Grant, P., & Basye, D. (2014). *Personalized learning: A guide for engaging students with technology*. International Society for Technology in Education.
- Hoffman, M., & Ragusa, C. (2015). Unwrapping GIFT: A primer on authoring tools for the Generalized Intelligent Framework for Tutoring. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym2)* (p. 11).
- Imhof, C., Bergamin, P., & McGarrity, S. (2020). Implementation of adaptive learning systems: Current state and potential. *Online teaching and learning in higher education*, 93-115.
- Kim, R., Olfman, L., Ryan, T., & Eryilmaz, E. (2014). Leveraging a personalized system to improve self-directed learning in online educational environments. *Computers & Education*, 70, 150-160.
- Kompen, R. T., Edirisingha, P., Canaleta, X., Alsina, M., & Monguet, J. M. (2019). Personal learning Environments based on Web 2.0 services in higher education. *Telematics and informatics*, 38, 194-206.
- Legault, N. (n.d.). *3 Simple Steps to Create Branched E-Learning*. E-Learning Heroes. <https://community.articulate.com/articles/3-simple-steps-to-create-branched-e-learning>
- Liu, Z. Y., Lomovtseva, N., & Korobeynikova, E. (2020). Online learning platforms: Reconstructing modern higher education. *International Journal of Emerging Technologies in Learning (iJET)*, 15(13), 4-21.
- Marienko, M., Nosenko, Y., & Shyshkina, M. (2020). Personalization of learning using adaptive technologies and augmented reality. *arXiv preprint arXiv:2011.05802*.
- Martin, F., Chen, Y., Moore, R. L., & Westine, C. D. (2020). Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. *Educational Technology Research and Development*, 68, 1903-1929.
- Martin, N. A., & Martin, R. (2015). Would you watch it? Creating effective and engaging video tutorials. *Journal of Library & Information Services in Distance Learning*, 9(1-2), 40-56.
- Peng, H., Ma, S., & Spector, J. M. (2019). Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment. *Smart Learning Environments*, 6(1). doi:10.1186/s40561-019-0089-y
- Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, 26, 582-599.
- Ryoo, & Winkelmann, K. (Eds.). (2021). *Innovative Learning Environments in STEM Higher Education*. Springer Briefs in Statistics. <https://doi.org/10.1007/978-3-030-58948-6>
- Slim, X. (2021). *CloudBerry Explorer for Amazon S3*. Softonic. <https://cloudberry-explorer-for-amazon-s3.en.softonic.com>
- Sottolare, R., & Brawner, K. (2018). 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 Conference of the Intelligent Tutoring Systems (ITS) Conference*, Montreal, Quebec, Canada.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Sottolare, R. A., Graesser, A., Hu, X., & Holden, H. (Eds.). (2013). *Design recommendations for intelligent tutoring systems: Volume 1-learner modeling* (Vol. 1). US Army Research Laboratory.
- Sottolare, R., Brawner, K., Goldberg, B., & Holden, H. (2017). The generalized intelligent framework for tutoring (GIFT). In *Fundamental issues in defense training and simulation* (pp. 223-233). CRC Press.
- Sottolare, R., McGroarty, C., Ballinger, C., & Aris, T. (2023). Investigating the Effect of Realistic Agents on Team Learning in Adaptive Simulation-based Training Environments using GIFT. In *Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11)* (p. 31). US Army Combat Capabilities Development Command–Soldier Center.
- Spector, J. M. (2014). Conceptualizing the emerging field of smart learning environments. *Smart learning environments, 1*, 1-10.
- Spector, J. M. (2016). The potential of smart technologies for learning and instruction. *International Journal of Smart Technology and Learning, 1*(1), 21-32.
- Taylor, D. L., Yeung, M., & Bashet, A. Z. (2021). Personalized and adaptive learning. *Innovative learning environments in STEM higher education: Opportunities, Challenges, and Looking Forward*, 17-34.
- Tzouveli, P., Mylonas, P., & Kollias, S. (2008). An intelligent e-learning system based on learner profiling and learning resources adaptation. *Computers & Education, 51*(1), 224-238.
- Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. *Journal of research on technology in education, 52*(3), 235-252.

## ABOUT THE AUTHORS

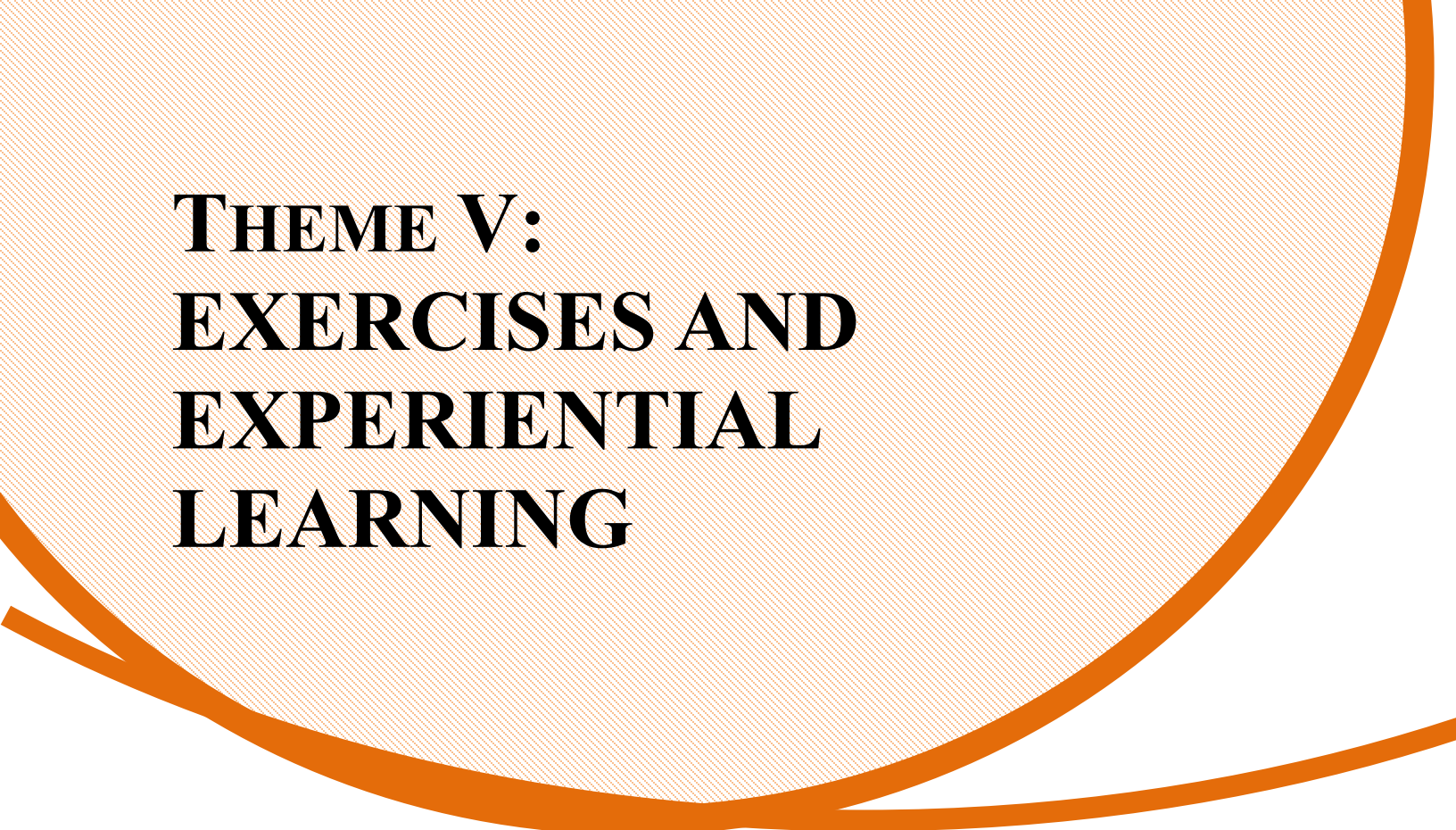
---

**Fadjimata I. Anaroua** is a Ph.D. student and research assistant at Embry-Riddle Aeronautical University. She was awarded a master's degree in software engineering at Embry-Riddle Aeronautical University, a bachelor's degree in Aeronautical Science, and a bachelor's degree in industrial engineering.

**Qing Li, Ph.D.** was awarded a Ph.D. in Learning Technologies from the University of North Texas in 2024. She is currently a lecturer at Chongqing University of Science and Technology. Her primary areas of research interest include technology-enhanced foreign language learning and teaching, as well as computer-aided translation and interpreting theory and practice.

**Hong Liu, Ph.D.** was awarded a Ph.D. in Mathematics and MS in Computer Science at the University of Arkansas, Fayetteville in 2000. He serves as a professor in Mathematics and Computing at Embry-Riddle Aeronautical University. He served as PI and Co-PI of 14 sponsored projects and published numerous articles in mathematics, data science, and STEM education. His current research interest is to develop a smart learning environment to promote peer learning.





**THEME V:  
EXERCISES AND  
EXPERIENTIAL  
LEARNING**





# STEEL-R in Multinational Joint Training Exercises (STEEL-Rx)

Aaron Presnall<sup>1</sup>, Biljana Presnall<sup>1</sup>, and Benjamin Goldberg<sup>2</sup>

Jefferson Institute<sup>1</sup>, U.S. Army Combat Capabilities Development Command (DEVCOM) – Soldier Center<sup>2</sup>

## INTRODUCTION

---

Operational integration of the Synthetic Training Environment Experiential Learning for Readiness (STEEL-R) system into large-scale multinational exercises promises to help mature STEEL-R under realistic network and scenario conditions. The goal is to enhance a Total Learning Architecture (TLA)-based Modular Open Systems Approach (MOSA) by strategically aligning performance evidence across active experiential learning events with competency frameworks that track learning across an ecosystem of resources (Goldberg et al., 2021). Additionally, these engagements will help facilitate interoperability among allies and partners and accelerate the associated training through personalization at scale and applied learning analytics. These outcomes, in turn, will support necessary advancements in force development for the U.S. and our allies.

In this paper we describe the early planning and progress for deploying STEEL-R in multinational joint training exercises. This project builds upon lessons learned from the implementation of learning analytics under the Office of the Secretary of Defense (OSD)-sponsored Maturing Advanced Distributed Learning in Exercises (MADLx) project, which collected empirical data from several events, including the VIKING, Combined Joint Staff Exercise (CJSE), and Bold Quest exercise series (Presnall & Radivojevic, 2018; Ljung et al., 2018; Salkutsan et al., 2021).

## Background

The STEEL-Rx project seeks to enhance military training and education by iteratively testing STEEL-R within a series of multinational joint training exercises. These interventions are expected to improve learning outcomes, increase learning efficiency, and expand readiness reporting through learning analytics and their visualizations in the exercises. The STEEL-R prototype will be tested in a series of exercise contexts (e.g., Bold Quest, Vigorous Warrior) and each of the targeted exercises will serve as a milestone to rapidly test, evaluate, and promote or discard solutions for core project objectives. Support for a given exercise includes its entire Joint Exercise Life Cycle, from planning to execution and through the final After Action Review (AAR). This work shall follow the following guidance:

- Department of Defense Directive 1322.18
- DoD Instruction 1322.26
- Joint Operational Training Infrastructure (JOTI) strategy
- CJCSM 3500.03E
- MC 458/2, NATO Education Training Exercise and Evaluation (ETEE) Policy
- Bi-SC Directives (75-2, 75-3, 75-7)

Large multinational civil-military exercises increasingly aim to leverage learning analytics as an integral part of the exercise experience. Prior work conducted in Bold Quest 21 and VIKING 22 (under the MADLx project) demonstrated the viability of this integration. Analytics on an aggregated dashboard with data from both the learning management system (LMS) and a Command-and-Control (C2) system allowed stakeholders to compare and correlate pre-training data to human performance data from the execution phase of the exercise (Presnall, 2020). These analytics were enhanced with data collected by designated evaluation teams and exit interviews with participants and exercise planners. Furthermore, diversified eLearning content with xAPI (experience API) reporting enabled show the impact to team performance in NATO brigade level exercises (Salkutsan et al, 2021).

STEEL-Rx would advance these efforts to a new level with novel methodologies. STEEL-R leverages tools and methods from the Army's Generalized Intelligent Framework for Tutoring (GIFT) in concert with TLA standards and projects (Goldberg et al., 2021). This led to a common data interoperability layer that collects evidence through a competency-based experiential learning model and functions across an ecosystem of distributed, synthetic, and live learning environments (Hernandez et al, 2022). This extension to the TLA provides data traceability and supports evidence-based training decisions through granular formative and summative assessments captured during an experiential scenario. STEEL-Rx aims to validate STEEL-R as an emerging capability to support adaptive learning, enable learning analytics, and deliver actionable results across complex, multi-platform, asynchronous learning and performance at scale. TLA standards and business practices will be applied to communicate outcomes to a competency management system for readiness and talent tracking and to a persistent data lake to support decision analytic pipelines.

Multinational exercises present a unique and rigorous testing environment for STEEL-R due to their vast scale, secure networking, diverse participants, and the frequent incorporation of legacy systems. These exercises also pose significant challenges, including cybersecurity requirements and intricate development timelines. However, their complex nature makes them ideal 'hard case' scenarios for evaluating the robustness of STEEL-R. Integration of STEEL-R in such a demanding context, along with its accompanying adaptive training and learning analytics, would demonstrate its capability and reliability. A successful deployment would strongly suggest that STEEL-R is versatile and robust enough to be effectively used in many operational settings.

### **Multinational Joint Exercises**

While no two exercises are alike, NATO member and partner nations generally comply with NATO directive Bi-SC CT-ED 75-3 for all major multinational distributed exercises, which organizes planning and execution processes into six steps: strategic guidance, design, prepare, execute, evaluate, and analyze. Each of the six sequential steps in the planning and execution cycle informs the subsequent step, as illustrated in Figure 1. In the U.S. Army, this is called the Joint Exercise Life Cycle (JELC) and anticipates a 440-day operations process. During the JELC, planners and staff hold regular and ad hoc meetings with one another, and with a wide range of stakeholders and subject matter experts. Key milestone planning conferences generally include a Concept Development Conference, Initial Planning Conference, Main Planning Conference, and Final Planning Conference. These gatherings are mainly to assure that exercise objectives, training objectives, the training path, and method for assessment of the participating units' execution of mission essential tasks are aligned within the complex array of organizational, logistical, and sometimes political factors inherent in multinational mission rehearsals.

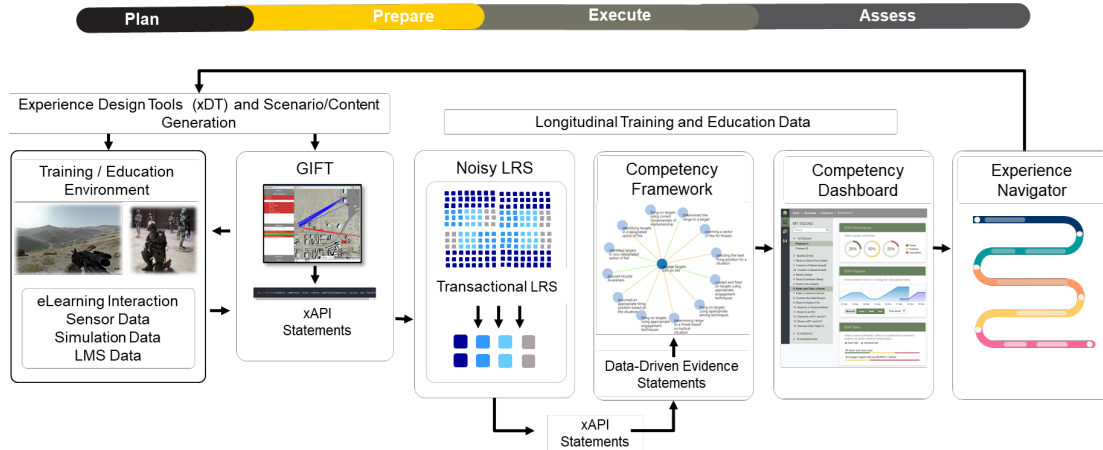


**Figure 1. Exercise Process**

The STEEL-Rx research findings are expected to provide a useful blueprint for military training stakeholders to indicate best practices for integrating supplementary learning and data-driven assessment into military exercises and for empowering observers/trainers with learning analytics visualizations. The practical implementation of TLA data and API standards will directly support ongoing work by the IEEE Learning Technologies Standards Committee (LTSC) and IEEE International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) to mature these standards—including xAPI, xAPI Profiles, Sharable Competency Definitions, and Competency Framework best practices—and learning engineering guidance associated with scenario and experience design. Thus, STEEL-Rx not only can enhance exercises but also advance science and technology in generalizable ways.

STEEL-Rx data collection from multiple sources will help ensure a broad range of insight. Each exercise can provide a summative testing opportunity, but additional formative testing will also be performed between exercises with relevant stakeholders as needed. Sources of data are expected to include:

- Data from built-in reporting tools in the exercise learning platforms (e.g., LMS, GIFT)
- Data from the exercise management system (CAX)
- Interviews with exercise planners, participants (training audience), observer/trainers, commanders
- Reaction surveys collected from stakeholders
- Pre- and post-tests on content, if needed, to supplement the automated data
- Results from the Exercise Evaluation (EXEVAL) and associated reports



**Figure 2. Notional Sequence for Operational Integration of STEEL-R in an Exercise**

See Figure 2 for a notional sequence for operational integration of STEEL-R in an Exercise. Design and development of capabilities for the selected joint and coalition exercises are anticipated to encompass instructional content (i.e., eLearning, virtual scenarios, live scenarios), competency frameworks, assessment materials, and research apparatus. Specific objectives include:

- Field-based validation testing of STEEL-R with exercise planners, observer/trainers, and trainees
- Support incorporation of assessment thinking into exercise planning and scenario design
- Empower commanders with intuitive, relevant analytics to inform decision-making
- Broaden the range and availability of demand-driven data and analytics in military exercises
- Improve the quality and utility of performance assessment (learning analytics) in exercises
- Support adoption of interoperable learning analytics standards and best practices (e.g., xAPI)
- Publish results in peer-reviewed scholarly venues to support advancement in this field

The anticipated outcomes of the STEEL-Rx project are manifold. It is expected to demonstrably improve the quality and utility of performance assessment in military exercises, expand the availability of learning analytics, and incrementally integrate these insights across the JELC. The STEEL-Rx project's iterative, evidence-based approach paves the way for the operational integration of the STEEL-R data strategy while identifying and validating best practices for the use of technology-enhanced learning in complex, dynamic, multinational training environments. The first two targeted exercises are Vigorous Warrior 24 and Bold Quest 24.

### **Vigorous Warrior**

Vigorous Warrior (VW) is the NATO Military Medicine Center of Excellence's (MILMEDCOE) biennial exercise to enhance the interoperability within the medical community focusing on the Supreme Allied Commander Europe's Guidance on Education Training Exercises and Evaluation (ETEE) and rooted in the Lessons Learned and Lessons Identified (LL/LI) of previous VW iterations. The training audience is about 1,200 persons from 39 nations. VW 24 will be executed April 22 - May 11 at the Bakony Combat Training Center in Hungary.

VW24 is a multinational, joint, multilevel medical LIVEX that will train the medical military structure in planning, coordinating and conducting medical functions during a NATO Major Joint Operation (MJO) in an Article 5 scenario.<sup>2</sup> The VW Aim is to enhance its effectiveness and interoperability of the military medical support system in a complex NATO MJO operating environment. VW explicitly seeks to serve as a platform for experimentation to test, tailor and further develop related concepts and capabilities.

STEEL-Rx engagement in VW24 will be an initial trust-building effort, with some modest substantive efforts, including: (1) assistance with the design and development of pretraining e-learning aligned with exercise and training objectives carefully designed to bring exercise participants to the same baseline knowledge level; (2) assistance with collection of standards-based learning analytics data; and (3) visualization of these analytics in an exercise dashboard for commanders to inform their decision making.

Our aim in 24 is to enhance the exercise experience, broaden the range and availability of demand-driven data and analytics, and support the delivery of pre-exercise academics using contemporary digital learning technologies and best practices. In doing so, we will seek to lay the foundations for evidence-driven training and exercise assessment as well as to begin to model a data-driven training mindset throughout the exercise lifecycle. Importantly, STEEL-Rx representation was invited to join the VW24 Exercise Evaluation (EXEVAL) team, and we are also working closely with the parallel Lessons Learned team, positioning us well to help shape the assessment thinking behind design and planning for VW26. We anticipate building on VW24 and the experience of Bold Quest 24 and 25 for a significantly more robust engagement in VW26 that takes advantage of multi-modal assessment across the exercise's experiential learning scenarios. This will provide a full implementation of STEEL-Rx that integrates evidence-centered data across the distributed learning and active learning opportunities designed into the training strategy and supporting scenarios.

### **Bold Quest**

Bold Quest 24 execution will be October 21 - November 6 at Camp Lejeune, North Carolina. STEEL-Rx's primary engagement will be within the Medical Thread, which shows promise as a relatively more permissive environment for operational integration of assessment thinking into planning and execution.

The Coalition Capability Demonstration and Assessment series, known as Bold Quest, is owned by the Joint Staff J6 Joint Fires Integration Division. Bold Quest was originally conceived in 2001 as an Advanced Concept Technology Demonstration (ACTD), with the first operational demonstration in 2003. The ACTD was extended twice at the request of the participant nations and services to accommodate an expanding scope of work. Bold Quest then transitioned from an ACTD to a recurring annual cycle of collaborative capability demonstrations and analysis.

It provides a platform for Joint and Coalition resource pooling, collaborative data collection, and data analysis to inform capability development on a Joint and Coalition scale. Interoperability across systems, services and nations is the central unifying aim of the BQ series across all threads. As an annual event, it allows for an iterative approach to test and develop capabilities. Largely centered around JADC2, BQ

---

<sup>2</sup> Article 5 of the North Atlantic Treaty, NATO's founding treaty, provides that an act of violence against any member of the NATO Alliance is to be considered as an armed attack against all members and all members are to assist the attacked Ally with all means necessary.

ultimately looks to support rapid and accurate information exchange, providing the warfighter with battlefield situational awareness to support decision making against modern and traditional opponents and increasing lethality among joint and coalition operations.

STEEL-Rx is looking to BQ24 as a venue to capture and translate large multi-modal LVC (live, virtual, and constructive) training event data sources to assist in the maturation of its competency-based data strategy for performance tracking, with specific attention to a complex multi stakeholder environment, where operational requirements are never far from mind. Our aim is to incrementally build upon the current STEEL-R infrastructure, including: (1) extensions to GIFT's domain modeling and xAPI capabilities, leveraging DoD (Department of Defense) open-source investments for automated individual and team assessment; (2) creating meaningful automated feedback and AAR interaction; (3) supporting intelligent, adaptive training at the scenario implementation level; and (4) driving evidence-based competency modeling and tracking across multiple learning resources and environments. We will leverage capabilities applied across medical training scenarios to measure and assess core competencies at the individual and team level.

To meet project objectives, we require access to various data sources/platforms/APIs to enable real-time data capture. A likely path will be focused on bridging human performance data on live and simulated use of BATDOK<sup>3</sup> in a Role 2 (field hospital) environment. We anticipate supplemental data from instrumented medical manikins and worn sensors on the medics to capture physiological indicators of stress and fatigue. In BQ 25, we hope to add a number of additional sensors, including cameras/mics for additional context within the multi-modal paradigm. An exciting added value is that BQ will enable GIFT to sit within the Mission Network, representing the first instance of GIFT within a JADC2 (Joint All Domain Command and Control) capability demonstration.

## **CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH**

This project will push the boundaries of STEEL-R as an emerging capability to support adaptive learning, enable learning analytics, and deliver actionable results across complex multi-platform asynchronous learning and performance at scale.

Several portions of this project can potentially transition into future use. First, field-testing will help mature STEEL-R directly and will provide actionable feedback to the larger PEO STRI program, particularly for the STE Training Management Tools. For example, this work will address goals identified by the TRADOC Proponent Office for STE (TPO-STE) to standardize data management requirements for robust analytics. This initially benefits DEVCOM and directly supports stakeholders such as Army University and its associated offices, including The Army Distributed Learning Program, the Army Training Information System (ATIS), member organizations associated with the Defense ADL Advisory Committee (DADLAC), and the Joint Staff J7 (Joint Training) including Joint Knowledge Online. Other collaborators such as the Defense Security Cooperation Agency (DSCA), NATO Allied Command Transformation (ACT), and

---

<sup>3</sup> *BATDOK is a tactical patient management platform that sits atop ATAK and was developed by AFRL.*

Partnership for Peace Consortium (PfPC) may also benefit from the direct products created through this work.

Second, the research findings from this project are expected to provide a useful blueprint for military training stakeholders; for instance, results are likely to indicate best practices for integrating supplementary learning into military exercises as well as ways to empower observer/trainers with learning analytics visualizations. These findings might inform, for example, supplements to the Joint Exercise Life Cycle (JELC), recommendations published by U.S. or NATO organizations such as the NATO *Advanced Distributed Learning Handbook*, or other guidance documents as the Government requests. Relatedly, the published research may advance some non-military training, such as first-responder exercises.

Lastly, the practical exercise of TLA data and API standards will directly support ongoing work by the IEEE LTSC and IEEE ICICLE to mature these standards, including xAPI, xAPI Profiles, Sharable Competency Definitions, and Competency Framework best practices, and learning engineering guidance. In other words, this work will advance science and technology in generalizable ways.

## REFERENCES

---

- Goldberg, B., Owens, K., Hellman, K., Robson, R., Blake-Plock, S., Hoffman, M. & Gupton, K. (2021). Forging proficiency and readiness through an experiential learning for readiness strategy. In Proceedings of the 2021 Interservice/Industry Training Simulation and Education Conference (I/ITSEC). Orlando, FL.
- Hernandez, M., Blake-Plock, Sh., Owens, K., Goldberg, B. & Robson, R. (2022). Enhancing Total Learning Architecture for Experiential Learning. In Proceedings of the 2022 Interservice/Industry Training Simulation and Education Conference (I/ITSEC). Orlando, FL.
- Ljung, M. N., Ax, M. T., Presnall, A., & Schatz, S. (2018). Integrating Advanced Distributed Learning into Multinational Exercises.
- Presnall, B. & Baker, S.R. (2020). Mapping eLearning Preparation to Training Objectives in Multinational Exercise: A Q-Matrix Approach. In Proceedings of the 2020 Interservice/Industry Training Simulation and Education Conference (I/ITSEC). Orlando, FL.
- Presnall, A. & Radivojevic, V. (2018). Learning Analytics with xAPI in a Multinational Military Exercise. In Proceedings of the 2018 Interservice/Industry Training Simulation and Education Conference (I/ITSEC). Orlando, FL.
- Salkutsan, S., Golovanov, A., Shuhyda, A., Tyschenko, M. & Presnall, B. (2021). Enhancing Military Exercise Team Performance with Diversified xAPI Instrumented eLearning. In Proceedings of the 2021 Interservice/Industry Training Simulation and Education Conference (I/ITSEC). Orlando, FL.

## ABOUT THE AUTHORS

---

**Aaron Presnall, Ph.D.** is president of the Jefferson Institute. Dr. Presnall serves as a Senior Advisor to the Office of the Secretary of Defense, Personnel and Readiness/Force Education and Training, and is National Chair of the NATO PfP Consortium Advanced Distributed Learning (ADL) Working Group, spearheading multinational cooperation to innovate eLearning and build interoperable, resilient, and agile training capabilities. As PI on the ADL Initiative MADLx project, he led the operational integration of eLearning and unified Learning and Performance Analytics in 11 multinational training events, providing commanders with data-driven insight for decision-making, and training management. He holds a Ph.D. in Politics from the University of Virginia.

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

***Biljana Presnall*** is Vice President of the Jefferson Institute with extensive experience in eLearning and military training. She led a digital team on a Department of Defense R&D project to mature the operational integration of Advanced Distributed Learning (ADL) in multinational exercises (MADLx) and contributed to the Annex to NATO ADL Handbook on ADL in Exercises. Her research is primarily focused on leveraging advanced data science methodologies within big data environments to develop innovative data strategies and solutions. This includes the application of artificial intelligence and natural language processing techniques to optimize and drive actionable insights.

***Benjamin Goldberg, Ph.D.*** is a senior research scientist at the U.S. Army Combat Capability Development Command – Soldier Center and is co-creator of the Generalized Intelligent Framework for Tutoring (GIFT). Dr. Goldberg is the team lead for a research program focused on the development and evaluation of Training Management Tools for future Army training systems. His research is focused on the application of intelligent tutoring and artificial intelligence techniques to build adaptive training programs that improve performance and accelerate mastery and readiness. Dr. Goldberg has researched adaptive instructional systems for the last 12 years and has been published across several high-impact proceedings. He holds a Ph.D. in Modeling & Simulation from the University of Central Florida.



# Automated Scenario Generation to Support Competency-Based Experiential Learning in GIFT

Andy Smith<sup>1</sup>, Randall Spain<sup>2</sup>, Wookhee Min<sup>1</sup>, Anne M. Sinatra<sup>2</sup>, Jonathan Rowe<sup>1</sup>, Bradford Mott<sup>1</sup>, and James Lester<sup>1</sup>

North Carolina State University, Center for Educational Informatics<sup>1</sup>, U.S. Army Combat Capabilities Development Command - Soldier Center<sup>2</sup>

## INTRODUCTION

---

Automated scenario generation is critical for meeting the Army's vision of providing unit leaders and Soldiers with engaging, relevant, novel synthetic training scenarios to support skill development and readiness (Goldberg et al., 2023). To effectively meet this goal, it is pivotal to explore how recent advancements in artificial intelligence (AI)-driven scenario generation can be leveraged. This technology holds the potential to automate the development of a diverse range of scenarios that address different competencies and contexts. Moreover, these scenarios can be tailored to the individual abilities of learners, thus ensuring a personalized learning experience. To meet this need, the U.S. Army Combat Capabilities Development Command Soldier Center, Simulation and Training Technology Center and North Carolina State University have launched a new collaborative effort to investigate how advances in AI-driven scenario generation can be used to automatically develop and deploy synthetic training scenarios to support competency-based experiential learning (CBEL). A key objective is to integrate these scenario generation functionalities into the Generalized Intelligent Framework for Tutoring (GIFT) and demonstrate these capabilities in the Synthetic Training Environment Experiential Learning for Readiness (STEEL-R) framework.

In this paper, we first seek to define the scenario generation problem for CBEL, including the needs of relevant stakeholders such as instructors and trainees, and the current non-automated scenario generation processes for developing synthetic training exercises. We investigate automated scenario generation through the lens of interactive narrative generation. Automated scenario generation and interactive narrative generation share key characteristics, including users who are active participants in virtual storyworlds that are dynamically personalized to users' actions and preferences, requiring scenarios that are realized in immersive simulation environments. Competency-based scenario generators are a form of interactive narrative generators that create adaptive narrative experiences, changing in response to learners' competency levels, as they solve problems and complete activities in synthetic training environments. In addition to the capabilities afforded by interactive narrative generators that can dynamically shape training experiences, story events, characters, and settings to enhance active learning and student agency (Wang et al., 2018), competency-based scenario generators can leverage recent advances in machine learning to further support trainees through competency-driven training.

Second, we present a competency-based scenario generation framework that includes three primary components, including (1) a scenario customization tool, with which instructors can enter requirements (e.g., commander's intent, learning objectives, target competencies) as a specification of the scenario being generated, (2) a scenario generator that takes as input instructor-provided requirements to produce a scenario representation (e.g., Extensible Training Service Package, XTSP) using AI-driven approaches (e.g., planner, large language models, deep neural networks), and (3) a scenario visualization and interaction tool, with which instructors and trainees can experience the scenario being generated. The scenario visualization and interaction tool takes into account human factors requirements revolving around trust in AI such as alignment with existing training scenarios, and inspectability and customizability of the scenarios through human-readable representations such as a concept of operations (CONOPs), warning orders (WARNO), and flow charts. In addition, the competency-based scenario generation framework can

incorporate a student model that utilizes assessment results from GIFT or external assessment engines and guides adaptive generation of scenarios based on dynamically changing learners' competency levels. The framework is designed to feature an iterative process of design, development, and refinement for creating effective competency-driven scenarios.

We conclude by highlighting promising approaches from the domains of narrative generation and procedural content generation, evaluating their strengths and limitations for competency-based scenario generation using Battle Drill 2A (BD2A) as a case study.

## RESEARCH CONTEXT

---

Many organizations, including the military, use simulation-based training to provide trainees with engaging, relevant, and novel training scenarios. When designed effectively, simulation-based training can be a powerful tool for applying skills trainees have learned in the classroom in a real-world context, thereby improving critical thinking (Chernikova et al., 2020), psychomotor learning (Rourke, 2020), teamwork (Laco & Stuart, 2022) and other technical (Brunye et al., 2020) and non-technical skills (Chae et al., 2021; Sinatra et al., 2021; Xie et al., 2021). Importantly, simulation-based training can be used to support CBEL by providing trainees rapid exposure to multiple scenarios within safe and controlled environments (Owens, 2021). CBEL prescribes an active approach to learning and expertise development that incorporates principles of guided experiential learning to provide learners with training sessions that are tailored to their own unique performance needs, competency states, and the inherent training needs. Affording learners the opportunity to apply and experience newly acquired skills is especially important during the early stages of learning (e.g., Crawl and Walk phases) when information retrieval and response generation processes are being formed and strengthened (Ericsson et al., 2018). A critical task in support of CBEL is competency-driven scenario generation, which describes how training scenarios should be designed, both from an instructional standpoint and fidelity perspective, to meet the learning needs of trainees.

Scenario generation for CBEL begins with defining the target competencies for a given learning experience, though competencies may be specified through many different representations, including mission essential task lists (METLs), or even specific training events related to competencies. In practice, this can take many forms depending on the level of the author chain. For example, a unit leader may provide a set of high-level tasks for a given training exercise, or perhaps only list the evaluation/certification events they would like their unit to prepare for. A scenario developer may then take this list and define a more granular set of tasks or training events for the scenario to support, and then convert that list to a playable scenario in a simulation environment, often drawing on a previously created training scenario. Existing tools like the Exercise Design Tool (EDT) have been designed to aid in creating live training exercises, while the Experience Design Tool (XDT) expanded the EDT to support the authoring of simulated training experiences (Mishra et al., 2023).

Promising approaches for automating aspects of the scenario generation process can be found from the research field of interactive narrative generation. Interactive narrative generation systems aim to take a defined setting, characters, and situation, and generate a storyline that adapts to player or character actions and authorial goals. Automated scenario generation and interactive narrative generation share several key characteristics, including active characters (both human and computer controlled) who are seeking to achieve some set of goals in a virtual storyworld environment. A variety of formalisms and computational techniques have been used to investigate these problems, including AI planners (Cavazza et al., 2002; Riedl & Young, 2010; Ware & Siler, 2021), reinforcement learning (Sawyer et al., 2017; Wang et al., 2018), and deep neural networks (Park et al., 2019).

## COMPGEN SCENARIO GENERATION FRAMEWORK

To address the challenges of scenario generation for CBEL, we propose the COMPGEN framework shown below in Figure 1. COMPGEN consists of three main components: (1) a scenario customization tool, (2) a scenario generator, and (3) a scenario visualization and interaction tool.

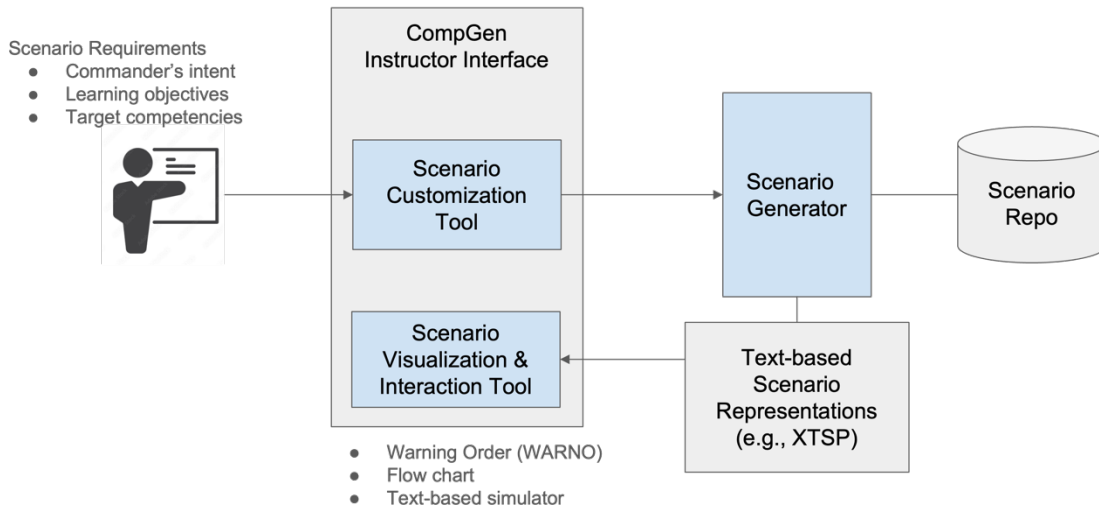


Figure 1. COMPGEN Scenario Generation Framework

As discussed above, the first step in the scenario generation process involves identifying the competencies, tasks, or training experiences that need to be included in the generated scenario. Building on lessons learned from previously developed systems like XDT, the scenario customization tool provides a set of user interfaces designed to allow a scenario author to communicate commander's intent and learning objectives in the form of a set of competencies and training events. This is also where scenario authors can specify other parameters related to the scenario, such as the target difficulty, stress level, and the incoming trainee or squad skill level.

The next phase of the pipeline is the scenario generator. The generator takes the set of competencies and tasks defined in the scenario customization tool, and generates a sequence of training events that provide opportunities for assessment of competencies. The scenario must also define assessments for each of the training events and generate the necessary configuration files for both the simulation environment and the assessment engine that will be collecting and analyzing user performance data. This output can take the form of actual configuration files or an intermediate representation such as XTSP (Hernandez et al., 2022), which describes mission parameters, actors, roles, teams, and events, to facilitate the integration and standardization of scenarios across various simulation environments.

Further tailoring of the generated scenario can then be completed using the Scenario Visualization and Interaction Tool. The goal of this module is to support visualizing the generated scenarios in multiple representations, such as CONOPs, WARNOs, or 3D visualizations in the simulation environment. This module differs from the Scenario Customization Tool, in that its focus is on more fine-grained changes to a given scenario, such as changing individual unit positions or types. After fine-grained changes have been made, the scenario can then be re-saved in the same representation as was generated by the Scenario Generator.

## AUTOMATED SCENARIO GENERATION

The most challenging component of the scenario generation framework above is the Scenario Generator. The Scenario Generator must take into account several factors as inputs, including things such as required competencies, learner ability levels, and targeted difficulty levels, while also accounting for the capabilities and restrictions of both the simulation environment and assessment engine that is generating the training scenario. In this section, we investigate three approaches to automated scenario generation, leveraging techniques from interactive narrative generation and procedural content generation.

### Narrative Planning

A popular computational technique for creating interactive narratives is to utilize automated or AI planners. Planners typically feature a world description, specified in a plan description language (e.g. PDDL or STRIPS) that describes the entities in the world and possible actions. Actions are defined with *preconditions* that dictate when the action can take place, and *effects* that describe how the action changes the state of the world. Given a set of initial conditions and a goal condition, the plan description can then be used by the planner to identify a valid sequence of actions to satisfy the goal condition. Though classical planners using only these features alone are insufficient for storytelling, researchers have expanded these systems and frameworks to create systems capable of incorporating more complex narrative features such as irony and suspense (Cardona-Rivera et al., 2024).

For CBEL, the plan definition can be tailored to align with training events in a typical scenario. For example, a BD2A scenario can be modeled to consist of three main phases. The squad must prepare and move to some location, the squad must respond to direct contact from enemy forces, and the squad must move to a location to exit the scenario. To generate such a scenario using a narrative planner, we must first author the plan definition file. For this example, we will be using the SABRE narrative planner (Ware & Siler, 2021).

```

/* Entities */
entity Squad : soldier;
entity OpFor : character;

entity A1 : location;
entity A2 : location;
entity A3 : location;
entity A4 : location;
entity A5 : location;
entity A6 : location;
entity A7 : location;

/* Properties */
property alive(character : character): boolean;
property at(character : character) : location;
property path(A : location, B : location) : boolean;
property AmbushCount() : number;

/* Initial State */
forall(c : character)
  alive(c);
at(Squad) = A2;
at(OpFor) = A7;
path(A1, A2) = True;
path(A2, A3) = True;
path(A3, A4) = True;
path(A3, A7) = True;
path(A4, A5) = True;
path(A4, A6) = True;

/* Actions */
action move(character : character, from : location, to : location) {
  precondition:
    from != to &
    alive(character) &
    at(character) == from &
    (path(from, to) == True | path(to, from) == True);
  effect:
    at(character) = to;
};

action Ambush(soldier : soldier, op : character, place : location){
  precondition:
    soldier != op &
    at(soldier) == at(op) &
    at(op) == place &
    at(soldier) == place;
  effect:
    AmbushCount() = AmbushCount() + 1;
};

/* Utilities */
utility():
  at(Squad) == A5 & AmbushCount() > 0;

utility(Squad):
  if (!alive(Squad))
    0
  else
    1;

```

Figure 2. Example plan definition file

The first step is to define the entities and properties of our world. Entities in this case consist of the Squad, and the opposition forces or OpFor. Entities also include the locations, or places of interest in the environment. Furthermore, properties can be assigned to Entities to describe their states. In this example,

properties include whether a given character is alive or dead, what location a character is currently at, if a path exists between two locations, and how many engagements have occurred in the current plan.

The next step is to define the initial state. As seen in Figure 2, the initial state for this scenario places the Squad at location A2, places the OpFor at A7, sets all characters to alive, and defines the valid paths between locations. There are two actions in this plan definition. The first action allows characters to move between locations that have a valid path between them. The Ambush action allows the OpFor to engage the Squad if they are at the same location. Finally, Utility (i.e., goal states) is defined. The *utility()* function defines authorial goals, in this example that the Squad makes it to location A5, and that they encounter at least 1 engagement. Utilities can also be defined for characters, in this case the Squad’s utility is defined by staying alive, though they could be expanded to incorporate more complex goal states leading to more intricate plans.

The SABRE planner uses the plan definition to generate a tree of potential actions and employs a feed-forward heuristic search to generate the plan shown in Figure 3. As depicted, both the Squad and OpFor move between locations using the paths defined, and one Ambush action occurs.

<b>Sequence of actions meeting the plan description outlined in Figure 2</b>
<code>move(Squad, A2, A3)</code>
<code>move(OpFor, A7, A3)</code>
<code>Ambush(Squad, OpFor, A3)</code>
<code>move(Squad, A3, A4)</code>
<code>move(Squad, A4, A5)</code>
<code>goal(AmbushCount &gt; 0 &amp; at(Squad) == A5)</code>

**Figure 3. Example plan generated by SABRE planner**

Parts of the plan definition, such as the initial state and utility, provide opportunities to customize the type of plans generated. The scenario developer can easily define different starting and end locations that would be updated in the initial condition and author/global utility of the example above. The author could also change how many Ambush encounters they would like the trainees to experience in a given scenario. The generated plan can then be post-processed into the appropriate configuration files for both a simulation environment and assessment engine. Locations can be converted into the appropriate coordinates for a 3D terrain, and the actions can be used to generate commands for computer controlled entities like the OpFor, as well as orders to be displayed for the Squad. Additionally, each action can be defined to enable the appropriate assessments for that action, as well as starting and ending triggers so that new assessments can be enabled as trainees progress through the scenario.

This example, while simple, showcases the planner's potential to generate multiple scenarios for a given world definition by reconfiguring the initial conditions and utility functions. Future work is needed to expand the set of actions to support more complex training scenarios, as well as to expand the complexity of the actions to better represent the phases of a training scenario and allow for a broader range of potential perturbations to vary difficulty and stress. Additionally, adaptations based on trainee ability and performance will need to be incorporated into the planning framework, either as expansions to the plan definition file, or as a director agent capable of interacting with both the simulation environment and planner in real time.

## Reinforcement Learning

Another promising approach to automated scenario generation is to formalize the problem as a reinforcement learning (RL) task by conceptualizing the scenario generator as an agent focused on adapting key dimensions of an exemplar training scenario to achieve instructor-specified objectives for training. Decisions about how to adapt different elements of a training scenario (e.g., terrain, unit location, unit behavior, time of day, mission objective) are each modeled as a Markov decision process (MDP). The MDP's state is encoded as a feature vector that summarizes the learner's current state, or in the case of dynamic scenario adaptation, the history of the learner's interaction with the generated scenario thus far. Actions represent the set of possible adaptations the generator can enact to augment a particular dimension of the exemplar scenario, as defined in the Scenario Adaptation Library (SAL). The reward function evaluates the desirability of an agent's actions, such as trainee performance or criteria based on instructor-specified goals (e.g., commander's intent, learning objectives, target competencies), which the scenario generator seeks to optimize. The solution to an RL-based scenario generation problem is a policy, or mapping between states and actions, that governs how the scenario generator produces new scenarios that differ from the selected exemplar scenario. For example, a separate SAL could be created for each phase of a BD2A scenario. The SALs would need to be created in such a way that every choice makes sense given previous adaptations, such as not having adaptations related to the squad leader, if the squad leader had been killed in a previous engagement. Then, at run-time different adaptations could be selected based on the squad's incoming ability level and performance up until that point in the scenario. RL provides a systematic process for exploring alternate approaches to automated scenario adaptation, gradually improving over time as more trainees interact with the scenario generator. In this BD2A example, this could involve identifying which combinations of adaptations produce the highest learning gains for different populations of trainees. Ideally, RL-based scenario generation models are induced using data from learner interactions with scenarios in a simulation-based training environment. However, synthetic data can also be utilized to bootstrap initial investigation into the particular formalization of a scenario generation model, including the state representation, action set, and reward model that have been chosen to formalize scenario generation decisions (Wang et al., 2018). See Figure 4 for the reinforcement learning-based scenario generation framework.

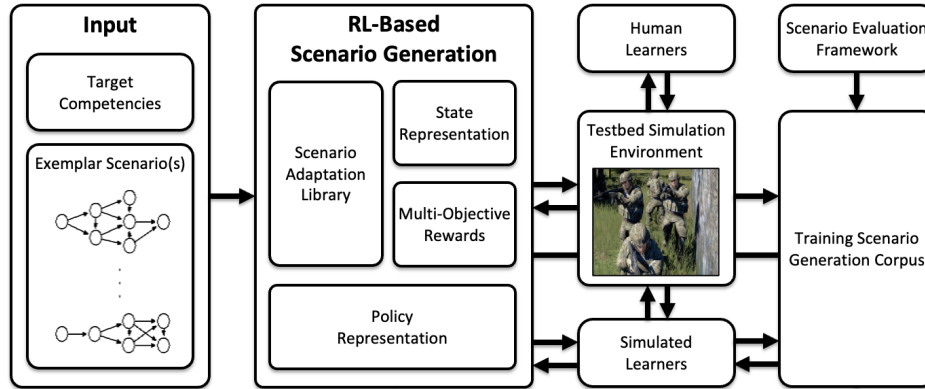


Figure 4. Reinforcement learning-based scenario generation framework

### Procedural Content Generation

Procedural content generation (PCG; Togelius et al., 2011) provides a framework for generating scenarios that meet instructor-specified objectives using machine learning models, while significantly reducing development costs. The project team has investigated PCG based on multistep deep convolutional generative adversarial networks (DCGANs), which demonstrated significant potential to create novel educational game levels (Park et al., 2019). This early work presented a multistep DCGAN framework, which takes iterative steps to generate novel, solvable levels. In the first iteration, a DCGAN is trained using a small set of solvable human-authored levels. Next, the DCGAN’s generator is used to create a large set of training examples, which are then filtered based on their solvability. These solvable levels are then used in combination with the human-authored levels as data augmentation to train another DCGAN generator with the objective of creating diverse, solvable levels. Through an iterative process following these steps, findings suggest that with only a small reduction in the novelty of the generated levels, the resulting generator exhibits significantly enhanced performance by generating a higher percentage of solvable levels compared to the generator trained only on human-authored levels. PCG can potentially create novel, adaptable scenarios that cater to the individual and collective needs of trainees. Generative modeling techniques, such as conditional generative adversarial networks, show significant promise for scenario generation tasks (C-GAN; Isola et al., 2017). For example, a C-GAN’s generator can be designed to take as input the competency levels of individual trainees and groups to generate a novel SAL that addresses diverse dimensions of trainees’ needs. Additionally, large language model-based approaches, promising for PCG using natural language interfaces, can design custom game levels and puzzles playable in an educational game (Kumaran et al., 2024).

## CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

Automated scenario generation is crucial for the success and continued advancement of competency-based experiential learning. Despite its importance, numerous questions regarding what techniques produce the best outcomes and how competency-based scenarios should be designed remain unanswered. A deeper understanding of the needs of instructors, scenario developers, and trainees will enable automated scenario generation systems to produce scenarios to effectively enhance learning outcomes. Additionally, converting these requirements into scenarios that can be embodied in 3D simulation-based training environments, as well as interfacing with existing assessment and tutoring frameworks like GIFT, is a labor-intensive and time-consuming task.

By leveraging research from the field of interactive narrative generation, we have identified several promising computational approaches to address this challenge. In the next year, we will continue to develop the library of adaptations and training events, and work with human experts to develop methods of empirically evaluating the performance of various automated scenario generation techniques in effectively supporting competency-based experiential learning. It will also be important to design, develop, and refine the different components of the scenario generation framework and integrate them into the larger Synthetic Training Environment (STE)/GIFT ecosystem.

## ACKNOWLEDGEMENTS

---

The research described herein has been sponsored by the U.S. Army DEVCOM, Soldier Center under cooperative agreement W912CG-19-2-0001. 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

---

- Brunyé, T. T., Brou, R., Doty, T. J., Gregory, F. D., Hussey, E. K., Lieberman, H. R., ... & Yu, A. B. (2020). A review of US Army research contributing to cognitive enhancement in military contexts. *Journal of Cognitive Enhancement, 4*, 453-468.
- Cavazza, M., Charles, F., & Mead, S. J. (2002). Character-based interactive storytelling. *IEEE Intelligent systems, 17*(4), 17-24.
- Cardona-Rivera, R. E., Jhala, A., Porteous, J., & Young, R. M. (2024). The Story So Far on Narrative Planning. In Proceedings of the 34th International Conference on Automated Planning and Scheduling.
- Chae, D., Yoo, J. Y., Kim, J., & Ryu, J. (2021). Effectiveness of virtual simulation to enhance cultural competence in pre-licensure and licensed health professionals: A systematic review. *Clinical Simulation in Nursing, 56*, 137-154.
- Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research, 90*(4), 499-541.
- Goldberg, B., Spain, R., Owens, K., Lanman, J., Kwon, C. P., Gupton, K., ... & Butler, P. A Data Strategy for Data-Driven Training Management: Artificial Intelligence and the Army's Synthetic Training Environment. In *Proceedings from the Interservice Industry Training Simulation and Education Conference*. NTSA
- Hernandez, M., Goldberg, B., Robson, R., Owens, K., Blake-Plock, S., Welch, T., & Ray, F. (2022). Enhancing the total learning architecture for experiential learning. In *Interservice/Industry Training, Simulation, and Education Conference (IITSEC)*.
- Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1125-1134). IEEE.
- Kumaran, V., Carpenter, D., Rowe, J., Mott, B., & Lester, J. (2024). Procedural Level Generation in Educational Games from Natural Language Instruction. *IEEE Transactions on Games*.
- Laco, R. B., & Stuart, W. P. (2022). Simulation-based training program to improve cardiopulmonary resuscitation and teamwork skills for the urgent care clinic staff. *Military Medicine, 187*(5-6), e764-e769.



## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

- Mishra, S., Owens, K., & Goldberg, B. (2023). The Evolution of an Experience Design Tool: From Live Exercises to Synthetic Experiential Learning with the Experience Design Tool and Integration with GIFT. In Proceedings of the eleventh annual GIFT users symposium (GIFTSym11) (pp. 74-86).
- Owens, K. P. (2021, July). Competency-based experiential-expertise and future adaptive learning systems. In International Conference on Human-Computer Interaction (pp. 93-109). Cham: Springer International Publishing.
- Park, K., Mott, B. W., Min, W., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2019, August). Generating educational game levels with multistep deep convolutional generative adversarial networks. In 2019 IEEE Conference on Games (CoG) (pp. 1-8). IEEE.
- Riedl, M. O., & Young, R. M. (2010). Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39, 217-268.
- Rourke, S. (2020). How does virtual reality simulation compare to simulated practice in the acquisition of clinical psychomotor skills for pre-registration student nurses? A systematic review. *International Journal of Nursing Studies*, 102, 103466.
- Sawyer, R., Rowe, J., & Lester, J. (2017). Balancing learning and engagement in game-based learning environments with multi-objective reinforcement learning. In Artificial Intelligence in Education: 18th International Conference, AIED 2017, Wuhan, China, June 28–July 1, 2017, Proceedings 18 (pp. 323-334). Springer International Publishing.
- Sinatra, A. M., Pollard, K. A., Files, B. T., Oiknine, A. H., Ericson, M., & Khooshabeh, P. (2021). Social fidelity in virtual agents: Impacts on presence and learning. *Computers in Human Behavior*, 114, 106562.
- Togelius, J., Yannakakis, G. N., Stanley, K. O., & Browne, C. (2011). Search-based procedural content generation: A taxonomy and survey. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(3), 172-186.
- 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 International Joint Conference on Artificial Intelligence (IJCAI)* (Vol. 18, pp. 13-19).
- Ware, S. G., & Siler, C. (2021, October). Sabre: A narrative planner supporting intention and deep theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Vol. 17, No. 1, pp. 99-106).
- Xie, B., Liu, H., Alghofaili, R., Zhang, Y., Jiang, Y., Lobo, F. D., ... & Yu, L. F. (2021). A review on virtual reality skill training applications. *Frontiers in Virtual Reality*, 2, 645153.

## ABOUT THE AUTHORS

---

**Andy Smith, Ph.D.** is a Research Scientist in the Center for Educational Informatics at North Carolina State University. He received his M.S and Ph.D. in Computer Science from North Carolina State University, and his B.S. degrees in Computer Science and Electrical and Computer engineering from Duke University. His research is focused on the intersection of artificial intelligence and education, with emphasis on user modeling, game-based learning, and educational data mining.

**Randall Spain, Ph.D.** is a Research Scientist in the Training and Simulation Division at U.S. Army Combat Capability Development Command – Soldier Center. He holds a Ph.D. in Human Factors Psychology from Old Dominion University. His research focuses on designing and investigating adaptive and intelligent training systems.

**Wookhee Min, Ph.D.** is a Research Scientist in the Center for Educational Informatics at North Carolina State University. He holds a Ph.D. in Computer Science from North Carolina State University. His research focuses on

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

*artificial intelligence centering on user modeling, natural language processing, educational data mining, and multimodal learning analytics.*

**Anne M. Sinatra, Ph.D.** is a Research Psychologist at U.S. Army Combat Capabilities Development Command Soldier Center, Simulation & Training and Technology Center in Orlando, FL. She has a Ph.D. in Applied Experimental and Human Factors Psychology from the University of Central Florida (UCF). Her research focuses on applying cognitive psychology and human factors principles to computer-based education and adaptive training to enhance learning. She is a member of the research team for the award winning Generalized Intelligent Framework for Tutoring (GIFT)

**Jonathan Rowe, Ph.D.** is a Senior Research Scientist in the Center for Educational Informatics at North Carolina State University, Adjunct Assistant Professor in the Department of Computer Science, and Managing Director of the National Science Foundation AI Institute for Engaged Learning. He earned his Ph.D. in Computer Science from North Carolina State University in 2013. His research focuses on designing, developing, and evaluating AI-augmented learning and training technologies, with an emphasis on game-based learning environments, multimodal learning analytics, interactive narrative generation, affective computing, user modeling, and intelligent tutoring systems.

**Bradford Mott, Ph.D.** is a Senior Research Scientist in the Center for Educational Informatics at North Carolina State University. He received his B.S., M.C.S., and Ph.D. in Computer Science from North Carolina State University. Prior to joining North Carolina State University, he worked in the video game industry developing cross-platform middleware solutions used extensively in commercial games and training applications. His research interests include AI and human-computer interaction, with applications in educational technology..

**James Lester, Ph.D.** is the Goodnight Distinguished University Professor in Artificial Intelligence and Machine Learning 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. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI).

# Implementing a Longitudinal Performance Comparison Interface for Improved After Action Review in Experiential Learning

Caleb Vatrak<sup>1</sup>, Naveeduddin Mohammed<sup>1</sup>, Nicholas Roberts<sup>3</sup>, Benjamin Goldberg<sup>2</sup>,  
and Gautam Biswas<sup>1</sup>

Vanderbilt University<sup>1</sup>, Dignitas Technologies<sup>2</sup>, US Army Combat Capabilities Development Command  
(DEVCOM) - Soldier Center<sup>3</sup>

## INTRODUCTION

---

Experiential learning stands as a cornerstone in modern educational methodologies, emphasizing the pivotal role of firsthand experiences in fostering deeper understanding and skill development (Andresen et al., 2020). Unlike traditional lecture-based approaches, experiential learning encourages learners to engage in hands-on activities, problem-solving tasks, and real-world applications to consolidate knowledge and enhance competency. Central to the effectiveness of experiential learning is the After Action Review (AAR) process, which provides a structured framework for reflection and improvement following learning experiences (Tokarieva et al., 2019).

During AAR, participants engage in a reflective discussion about strengths and areas for improvement. Under the direction of a trained instructor, this discussion often employs structured frameworks and techniques to guide the conversations and encourage in-depth reflection. One common component of these techniques is the use of comparative examples and contrasting cases, which provide directed feedback and examples for how trainees can improve their skills and strategies (Hanoun et al., 2018). Two primary types of performance comparisons are commonly employed: expert comparison and longitudinal past performance comparison. Expert comparison involves benchmarking against expert models to evaluate performance levels and problem-solving approaches against a “gold standard”, offering valuable insights for improvement. On the other hand, longitudinal past performance comparison entails assessing current performance against previous performance, enabling learners to track progress over time and identify areas for growth.

Despite the recognized benefits of such comparative analysis (Schwartz et al., 2011; Sidney et al., 2015), the use of such comparative methods has historically been difficult due to challenges with data management and user interface (Chromik & Butz, 2021; Vatrak et al., 2022a). Working toward solving these issues, previous research has demonstrated the implementation of expert comparison interfaces, such as the one integrated into the Generalized Intelligent Framework for Tutoring (GIFT) Game Master Dashboard (Sottolare, et al, 2017; Vatrak et al., 2023a). However, this implementation was limited to only expert comparison, thereby missing the key insights generated through longitudinal past performance comparison, hindering the full potential of comparative AAR experiences. Building upon this prior work, this paper introduces an extension to the Game Master playback dashboard interface, enabling longitudinal tracking of learner performance for comprehensive performance comparison. In this paper, we delve into the design considerations and implementation details of the prior performance comparison interface. Ultimately, we envision that such a comparative analysis tool will prove invaluable for learners and instructors alike, augmenting AAR practices and fostering continuous improvement in learning experiences.

## BACKGROUND AND RELATED WORK

---

### Experiential Learning and AAR

Experiential learning, rooted in the philosophy of learning by doing, emphasizes the active engagement of learners in practical experiences to facilitate knowledge acquisition and skill development (Andresen et al., 2020). This approach diverges from traditional lecture-based methods, instead advocating for hands-on activities, problem-solving tasks, and real-world applications to deepen understanding and enhance retention. One of the most influential models of experiential learning is David Kolb's theory (Kolb, 2014), which outlines a four-stage cycle:

1. **Concrete Experience:** This is the first stage where learners encounter a new experience or situation. It could be anything from a hands-on activity to a real-life problem. The experience is the starting point for learning.
2. **Reflective Observation:** After the experience, learners reflect on what happened. They consider their thoughts and feelings during the experience, trying to understand what occurred and why. Reflection is crucial for extracting meaning from the experience.
3. **Abstract Conceptualization:** In this stage, learners attempt to make sense of their observations by creating theories or concepts. They seek to understand the patterns underlying their experiences, forming generalizations and principles that can be applied in various contexts.
4. **Active Experimentation:** Finally, learners apply their newly formed concepts and theories to different situations. They actively experiment with what they have learned, testing their understanding and adapting it as necessary based on feedback and further experiences.

Kolb's model is most often depicted as a cycle because learning is seen as a continuous process where each stage informs the next. Individuals may enter the cycle at any stage and may revisit stages multiple times as they engage in the learning process. This model emphasizes the importance of both concrete experience and reflective observation in effective learning, as well as the iterative nature of learning through experimentation and conceptualization.

Central to practical implementation of the experiential learning paradigm is the AAR process, also sometime known as debriefing, which serves as a structured mechanism for reflection and improvement following learning experiences (Hanoun et al., 2018). AAR typically involves reviewing actions taken during an exercise, guided by domain experts or instructors, to identify successes, challenges, and areas for development. The AAR process is highly related to the *Reflective Observation* and *Abstract Conceptualization* components of Kolb's model, as it allows learners to reflect on their hands-on experiences and translate those experiences into meaningful learning outcomes. By fostering self-reflection and facilitating dialogue, AAR promotes continuous learning and skill refinement.

## Comparative Analysis within AAR

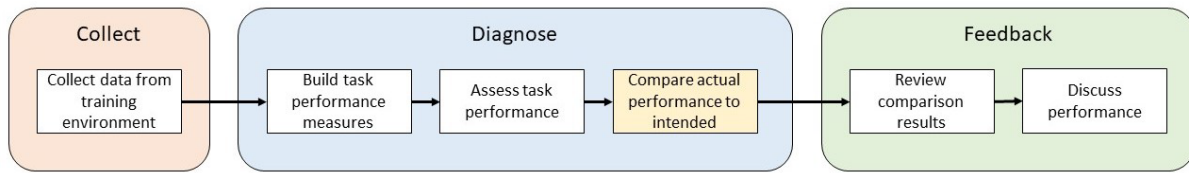


Figure 1. The three-phase AAR framework adapted from Hanoun et al. (2018).

Within the AAR process, there are many structured models and frameworks to help instructors facilitate effective reflection among trainees. Often within such models, one suggested component is a comparative analysis of performance. For example, Hanoun et al. (2018) proposed a theoretical model of AAR, which breaks down the process into three phases: (1) collection, (2) diagnosis, and (3) feedback, as depicted in Figure 1. In this model, the end component of the diagnosis phase is *comparison of actual performance to the intended performance*, representing this idea of a comparative analysis. Such methods are important as they provide trainees with a means to assess individual or team performance, understand how they have grown during training so far, and identify areas for continued improvement (Visscher & Coe, 2003). Within comparative analysis, there are two primary types commonly employed:

1. **Expert Comparison:** Involves benchmarking against expert models to evaluate performance levels and problem-solving approaches in reference to a “gold standard”. By comparing trainee performance to established standards of expertise, expert comparison facilitates the identification of strengths and weaknesses and adopts practices that have been proven effective and successful by experts in the field, offering valuable insights for skill development and enhancement.
2. **Longitudinal Past Performance Comparison:** Entails assessing current performance against previous performance, enabling learners to track progress over time and identify trends or patterns in skill development. By examining performance trajectories, longitudinal comparison empowers learners to recognize established growth, identify areas for improvement, and tailor learning strategies accordingly. By comparing present performance to previous performance, trainees and their instructors can assess the effectiveness of their program and decide on future training strategies.

Expert and historical performance comparison can be paired to offer an effective structured framework for AAR reflection, due to the distinct advantages of each kind of performance comparison during AAR. Organizations and trainees can gain a more thorough picture of their performance and can make well-informed judgments about their future training requirements by combining the two forms of comparison.

## Prior Work on Performance Comparison in GIFT

In previous work, we explored the development and implementation of toolsets within GIFT to streamline the performance comparison process (Vatral et al., 2023a). By implementing a flexible expert comparison interface within the GIFT Game Master dashboard, we were able to simplify the process of adding a performance comparison component to the AAR processes that utilize GIFT. This comparison interface was carefully designed to be easy to implement and extend across a wide variety of training drills, consistent with the design philosophy of the GIFT software. However, despite the advantages of this prior implementation, that work primarily focused on development of *expert* comparisons within Game Master, overlooking the advantages of tracking longitudinal performance over multiple drills and combining multiple types of comparative analysis. As such, there is a need for extensions to the existing Game Master

interfaces to enable longitudinal comparison, thereby providing a comprehensive view of learner progress and facilitating targeted feedback and improvement strategies. In the remainder of this paper, we describe such an implementation.

## DESIGN GOALS AND CONCEPTUALIZATION

---

In this section, we outline the design goals for the new longitudinal performance comparison interface. The design of the longitudinal performance comparison interface is guided by specific objectives aimed at enhancing the AAR process, drawing from both prior work in GIFT and best practice theory for comparative analysis. Among these objectives are the three critical design goals.

1. **Facilitating Comprehensive Performance Analysis:** The primary aim of the longitudinal performance comparison interface is to provide users with a detailed understanding of learner performance across multiple training sessions. The interface empowers instructors and learners to gain insights into performance trends, identify areas of consistency or variation, and discern patterns of improvement or stagnation over time. This holistic view facilitates a deeper understanding of individual and team performance dynamics, thereby fostering informed decision-making and targeted intervention strategies for continued training.
2. **Supporting Targeted Feedback and Improvement Strategies:** Central to the efficacy of the longitudinal performance comparison interface is its role in supporting targeted feedback and improvement strategies. By tracking performance progression over time, the interface equips instructors and learners with the tools to identify emerging trends, patterns, and areas for improvement. By allowing both high-level longitudinal analysis and low-level comparison on specific exercises, this analysis enables the formulation of targeted feedback and intervention strategies tailored to address specific performance gaps and enhance learning outcomes.
3. **Enhancing User Experience and Usability:** A pivotal aspect of the design process revolves around ensuring an optimal user experience and usability for all stakeholders involved in the AAR process. The interface is carefully designed to be intuitive, user-friendly, and conducive to dynamic AAR practices. Furthermore, the interface is designed to be flexible and extensible, accommodating diverse user preferences and evolving educational requirements. This balance between usability and flexibility ensures that the interface remains adaptable to the dynamic needs of end-users, while also offering developers and course authors the necessary tools and frameworks to extend its functionality and tailor it to specific educational contexts.

These design goals of performance analysis, targeted feedback, and usability, guided the design and development process of the new interface, consistent with both the philosophies of GIFT and theories for effective AAR. To realize these design goals, we conceptualized a two-part prior performance comparison interface. The first part of the new tools is designed as an interface for high-level longitudinal performance comparison, hereafter referred to as the *longitudinal view*. This allows users to create a color-coded visual representation of performance over the course of several exercises, inspired by the performance progression plots seen in Vatral et al. (2022b, 2023b). This gives the instructors and trainees a high-level overview of how performance within various metrics is progressing across multiple exercises, allowing them to gain high-level insights about the specific competencies that require attention and more practice. Thus, this longitudinal view allows for significant insights into training program design, for example, by selecting exercises that will maximize practice on competencies that have failed to increase under current training conditions.

The second part of the new tools is designed as an interface for detailed low-level comparison between two specific instances of a training drill, hereafter referred to as the *direct view*. This interface shows user graphical and textual elements across a variety of defined metrics that directly compare low-level performance and strategies between two runs of an exercise, directly inspired by the expert performance comparison interface seen in Vatrál et al. (2023a). In contrast to the longitudinal view, this low-level direct comparison provides specific insights into the behaviors and strategies that trainees have used and how they affect the measured performance levels. By combining insights from both the longitudinal view and direct view, instructors and trainees can understand performance trends and make plans for how to improve based on high- and low-level evidence.

## IMPLEMENTATION DETAILS

---

### GIFT and the Game Master Dashboard

Building upon prior work and existing toolsets, the new past performance comparison interface is built in GIFT atop the existing Game Master dashboard. The U.S. Army created the open-source GIFT software system to provide a comprehensive toolset for intelligent computer-assisted education and training programs (Sottolare et al., 2017). Learner interactions with the system and other assessments are among the many sources of data that GIFT employs to dynamically modify the educational material and delivery in order to maximize learning results. With its modular architecture, GIFT facilitates the creation of intelligent tutoring systems (ITSs) by offering a range of tools and modules that enable developers to customize the system to meet their unique requirements. The Game Master dashboard is one such tool that lets users manage and monitor the current game state and the events that are happening both during and after training sessions. Session replay, which streamlines debriefing and AAR, is one of Game Master's primary features. It lets instructors and trainees examine the specifics of and performance during a prior training session on a dynamic timeline interface. Because of its utility in supporting AAR processes, prior work has extended the Game Master tools to support a comparative analysis between expert data and trainee sessions as part of an AAR.

### Extending the Expert Comparison Interface

Building upon the foundation laid by the previously developed expert comparison interface in GIFT's Game Master Dashboard (Vatrál et al., 2023a), the implementation of our longitudinal prior performance comparison interface involves extending existing interfaces and APIs to enable prior performance comparison. Within this existing expert comparison implementation, performance comparison is defined via a Java interface, which condition classes simply implement to define their expert comparative analysis capabilities. Two parameters are sent into the comparison function defined in the interface: a JSON object that represents the defined expert model and the data from the current Game Master session. To provide for maximum flexibility, the condition class designers decide how to proceed with the comparison's remaining logic by implementing this function. The function's return signature is defined by a flexible XML schema managed by GIFT's AbstractSchemaHandler. The returned XML schema defines what should be displayed back to the user in Game Master for the comparison, based on several templates from which the assessment designer can select. There are currently two templates in the existing implementation. First, the single comparison template provides structure for a single multimedia element (image, video, chart, etc.) placed above a list of additional text elements. This might be applied, for instance, when the condition class generates a single chart or image to illustrate the comparison along with text descriptions of the image or additional numerical measurements. In contrast, the side-by-side comparison template shows a list of text items displayed below two multimedia elements, each with a designated title. This might be applied, for

instance, if the designer wants the user to make a direct comparison between a trainees' performance image and an expert's performance image. Figure 2 shows a diagram of these two comparison templates.

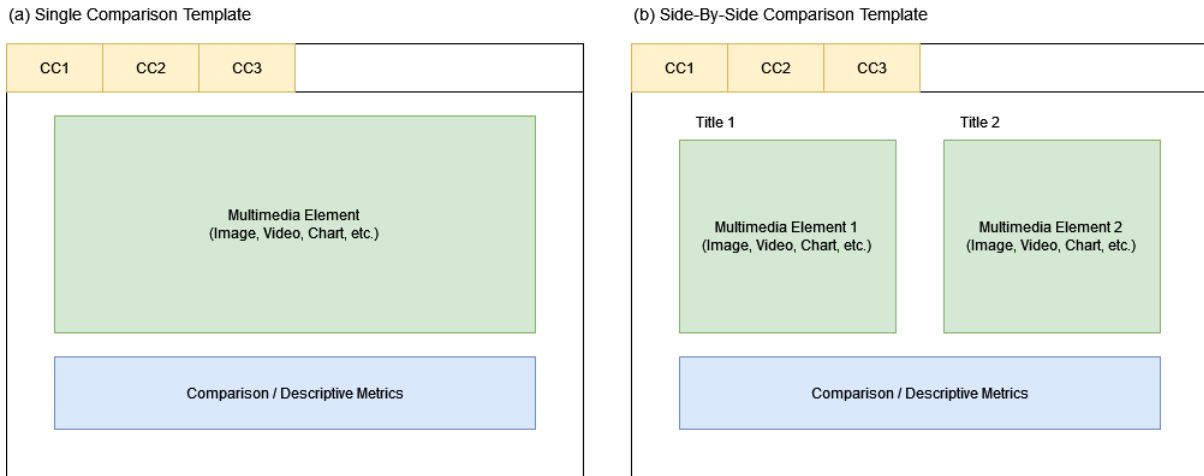


Figure 2. Block diagram of the implemented direct comparison dashboard templates.

The *direct view* of the prior performance comparison implementation, which focuses on detailed performance comparison between two specific drills, heavily derives from and extends this expert comparison functionality. At a high-level, the details and data flow of the prior performance direct view are almost identical to the expert comparison view. The end user opens a session in Game Master, selects prior performance from the experience analysis menu and chooses a prior session to compare to the session currently open in Game Master. At this point, Game Master compares which condition class assessments are common between the two sessions and sends both the current and prior data to these condition classes for processing, which ultimately returns the data needed to display this direct comparison view. At an API level, the implementation is also highly similar. We extended the experience analysis interface previously implemented for expert comparison with an additional method designed to perform the prior comparison

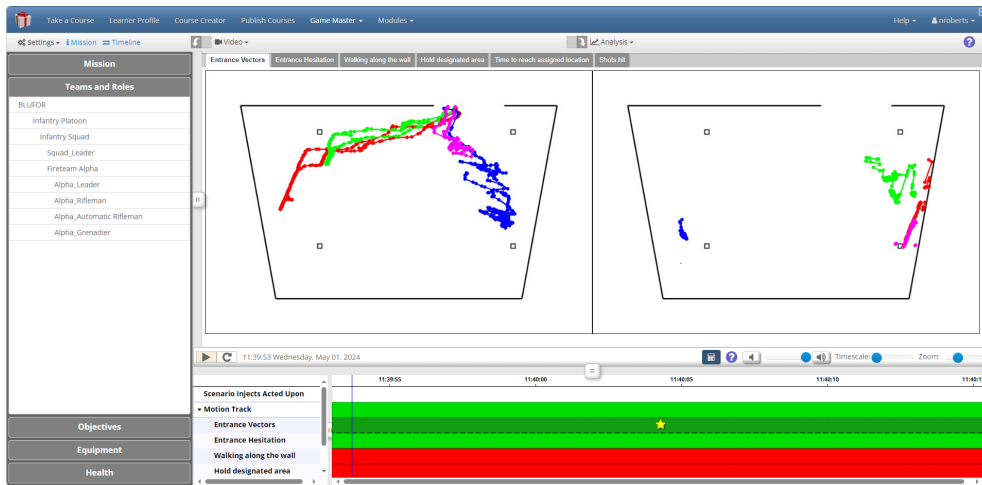


Figure 3. Example of the direct prior performance comparison view using the side-by-side template.



for each condition class. The function signature is also almost identical, with the primary difference being that the new prior comparison function takes in two sets of data from Game Master, rather than one set of data and one JSON expert model. The return signature uses the same extensible XML schema to define the graphical and textual elements to be displayed in the Game Master interface.

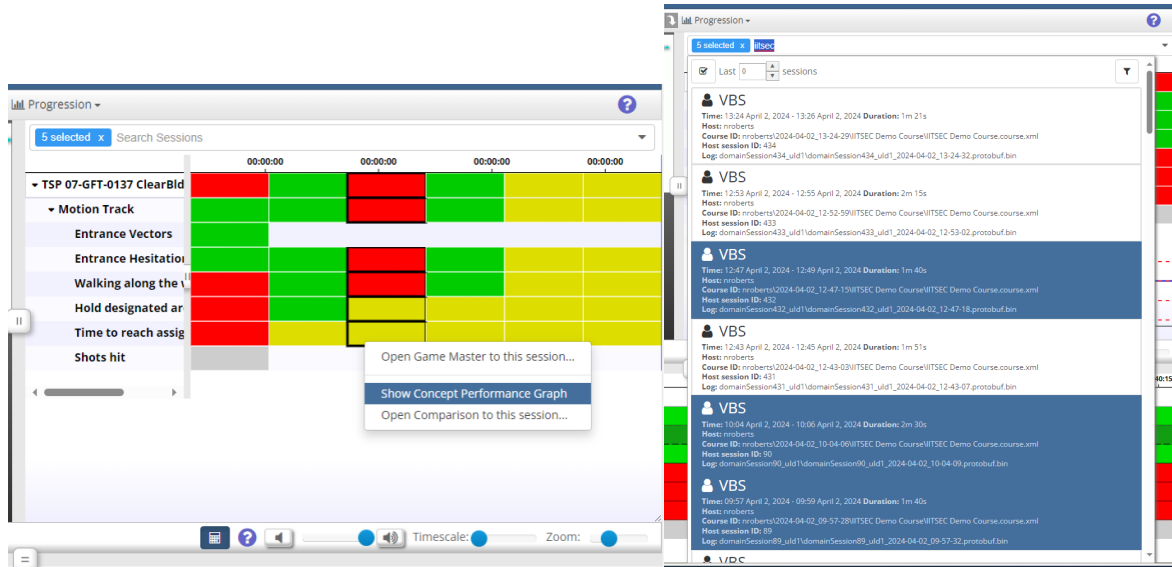


Figure 3. Longitudinal performance comparison interface. Left shows the dynamically generated plot of performance. Right shows the interface to multi-select sessions to be included in the plot.

### Longitudinal Performance Plot

The *longitudinal view* of the prior performance comparison implementation, which focuses on high-level performance comparison between any number of prior sessions, was implemented as a new feature in this work. The overall user workflow is similar to other components of the experience analysis feature set but utilizes a different set of data processing and graphical display features internally. The end user opens a session in Game Master and selects longitudinal performance from the experience analysis menu. At this point, a new modal dialogue window pops up and populates with the previously recorded sessions, allowing the user to multi-select the sessions that they wish to include in the longitudinal comparison, as shown in Figure 4 (Right). Some basic filtering options – such as date, time, course name, etc. – are built into this selection window in the prototype to more easily allow users to select only relevant past sessions, but more advanced filtering, such as filtering based on assessments included in the session, will be added in future work. After confirming their selection, Game Master generates and displays a color-coded plot of performance on each included assessment across time, as shown in Figure 4 (Left). This plot is interactive, allowing users to click on individual elements to see more information about the session and recorded metrics or to optionally open the direct view to a comparison between the active session and the selected session, allowing quick access to a more in-depth analysis. At a technical level, this plot is generated by querying the summative scores for each active assessment in each of the selected sessions. However, this initially created a problem where some assessments did not have summative scores, as they only recorded formative values. To fix this issue and facilitate consistency across assessments in this interface, a new rule type was introduced to the domain knowledge file to populate summative assessment scores based on the last received value for a formative assessment. This is the default behavior if another summative assessment rule is not implemented.

## **CONCLUSION AND RECOMMENDATIONS FOR FUTURE RESEARCH**

---

In this paper, we have presented an ongoing project to extend the comparative experience analysis features in GIFT by developing a longitudinal prior performance comparison interface to support AAR. The new features contain two primary interfaces. First, the *longitudinal view* allows instructors and trainees to see a dynamically generated color-coded timeline of performance across multiple assessment metrics. This high-level view of performance progression allows stakeholders to analyze trends in performance across multiple executions of an exercise in a training course. This has significant potential to impact training strategy modification, both during the execution of a training program through rapid adjustments and tailoring the lessons to individual trainee needs, as well as after a program is over through review of trainee learning results and modification of program design accordingly. In addition, this longitudinal comparison view can be utilized with very little modification to existing courses and their domain knowledge files, allowing for immediacy of its impact. Second, the *direct view* allows instructors and trainees to see in-depth, detailed comparisons between two specific training sessions. This low-level view of performance progression allows stakeholders to analyze specific strategy differences between two exercises and how those changes affected the measured performance. This has significant potential to impact day-to-day and exercise-to-exercise performance of trainees by allowing them to quickly test new skills and strategies and directly compare their impact on performance. This direct comparison interface is designed to be extremely flexible for course designers, to allow them to show trainees detailed feedback that is tailored to their specific performance outcome measures. However, despite this flexibility, the direct comparison remains easy to implement by building on top of the existing condition class structure within GIFT. By utilizing both new comparison views, instructors, trainees, and course designers can more holistically understand performance and the efficacy of a training program by promoting in-depth reflection at both a low-level operational viewpoint and a high-level longitudinal viewpoint.

There are several promising areas for future work gleaned from this ongoing project, both in the specific development of GIFT's comparative experience analysis tools, as well as for the larger ITS development community. One of the critical next steps for this work is testing the comparative experience analysis interface's usability and efficacy with a variety of stakeholders, such as instructors, trainees, and subject matter experts. This will involve gathering input on the interface design, performing user studies, and evaluating the interface's effect on learning objectives. Not only would these studies benefit the ITS design community by providing design suggestions for the interface within GIFT or for similar comparative interfaces in other software systems, but it would also benefit the simulation-based training community at large by providing more empirical evidence for best practices within debriefing and AAR. In addition, future development work should focus on the creation of authoring tools in GIFT to allow for easier design and implementation of these experience analysis features into existing courses. Where right now, implementation of the direct view involves modifying the codebase of condition classes, graphical authoring tools would significantly lower the barrier to entry and provide at least some basic version of these tools to a wider audience. The implementation of such graphical authoring tools presents significant challenges for future work both in user design and technical implementation.

With continued development and testing, we believe that comparative experience analysis tools such as those described in this paper have significant potential to be highly valuable for improving AAR in a wide variety of experiential learning contexts. As we look toward the future, it is important to continue research and innovation in educational technologies that can support and improve the educational outcomes for the next generation of learners.

## REFERENCES

---

- Andresen, L., Boud, D., & Cohen, R. (2020). Experience-based learning. In *Understanding adult education and training* (pp. 225-239). Routledge.
- Chromik, M., & Butz, A. (2021). Human-xai interaction: A review and design principles for explanation user interfaces. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18* (pp. 619-640). Springer International Publishing.
- Hanoun, S., & Nahavandi, S. (2018). Current and future methodologies of after action review in simulation-based training. *2018 Annual IEEE International Systems Conference (SysCon)*, 1–6.
- Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press.
- Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of educational psychology*, *103*(4), 759.
- Sidney, P. G., Hattikudur, S., & Alibali, M. W. (2015). How do contrasting cases and self-explanation promote learning? Evidence from fraction division. *Learning and Instruction*, *40*, 29-38.
- Sottolare, 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*, 1-19.
- Tokarieva, A. V., Volkova, N. P., & Harkusha, I. V. (2019). Educational digital games: models and implementation.
- Vatral, C., Mohammed, N., & Biswas, G. (2022a). Promoting Explainable Feedback in Simulation-Based Training through Contrasting Case Exemplars. In *Virtual Workshop Proceedings: Advances and Opportunities in Team Tutoring* (p. 11).
- Vatral, C., Biswas, G., Mohammed, N., & Goldberg, B. S. (2022b). Automated Assessment of Team Performance Using Multimodal Bayesian Learning Analytics. In *Proceedings of the 2022 Interservice/Industry Training, Simulation and Education Conference (IITSEC)*. National Training and Simulation Association.
- Vatral, C., Mohammed, N., Biswas, G., Roberts, N., & Goldberg, B. (2023a). A Comparative Analysis Interface to Streamline After-Action Review in Experiential Learning Environments. In *Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11)*. US Army Combat Capabilities Development Command–Soldier Center.
- Vatral, C., Mohammed, N., Goldberg, B. S., & Biswas, G. (2023b). A Framework for Performance Assessment Across Multiple Training Scenarios Using Hierarchical Bayesian Competency Models. In *Proceedings of the 2023 Interservice/Industry Training, Simulation and Education Conference (IITSEC)*. National Training and Simulation Association.
- Visscher, A. J., & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School effectiveness and school improvement*, *14*(3), 321-349.

## ABOUT THE AUTHORS

---

**Caleb Vatral** is a Ph.D. candidate at Vanderbilt University in the Department of Computer Science. Drawing from theories of distributed cognition and experiential learning science and combining them with research methodologies from user-centered design and computational techniques from multimodal learning analytics, Caleb's research focuses on how to improve education and training through the integration of AI technologies. By focusing on the strengths of both humans and AI systems, how these strengths can complement one another to build new capabilities, and how these capabilities can be designed and integrated into teaching with a strong emphasis on stakeholder needs and desires, Caleb's work aims to build sustainable human-AI teaming in educational contexts.

**Naveeduddin Mohammed** is a Senior Research Engineer with the Institute for Software Integrated Systems at Vanderbilt University. Naveed received the M.S. degree in Computer and Information Sciences from University of

## Proceedings of the 12th Annual GIFT Users Symposium (GIFTSym12)

*Colorado. He is a full stack developer, and his work focuses on designing, developing, and maintaining frameworks for open-ended computer-based learning environments and metacognitive tutors.*

***Nicholas Roberts** is a senior software engineer at Dignitas Technologies and the engineering lead for the GIFT project. Nick has been involved in the engineering of GIFT and supported collaboration and research with the intelligent tutoring system (ITS) community for nearly 10 years. Nicholas contributes to the GIFT community by maintaining the GIFT portal ([www.GIFTTutoring.org](http://www.GIFTTutoring.org)) and GIFT Cloud ([cloud.gifttutoring.org](http://cloud.gifttutoring.org)), supporting conferences such as the GIFT Symposium, and technical exchanges with Soldier Center and their contractors. He has also been heavily involved in integrating GIFT into TSS/TMT.*

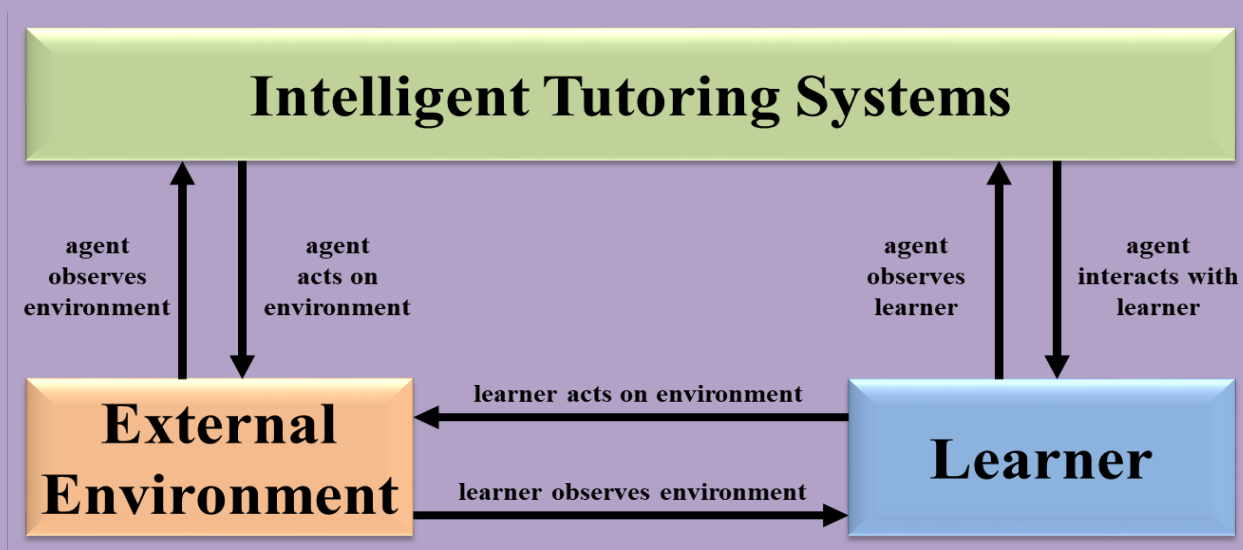
***Benjamin Goldberg, Ph.D.** is a senior research scientist at the U.S. Army Combat Capability Development Command – Soldier Center, and is co-creator of the Generalized Intelligent Framework for Tutoring (GIFT). Dr. Goldberg is the technical lead for a research program focused on the development and evaluation of Training Management Tools for future Army training systems. His research is focused on the application of intelligent tutoring and artificial intelligence techniques to build adaptive training programs that improve performance and accelerate mastery and readiness. Dr. Goldberg has researched adaptive instructional systems for the last 15 years and has been published across several high-impact proceedings. He holds a Ph.D. in Modeling & Simulation from the University of Central Florida.*

***Gautam Biswas, Ph.D.** is a Cornelius Vanderbilt Professor of Engineering and Professor of Computer Science and Computer Engineering at Vanderbilt University. He conducts research in Intelligent Systems with primary interests in monitoring, control, and fault adaptivity of complex cyber physical systems, as well as developing intelligent open-ended learning environments that adapt to students' learning performance and behaviors. He has developed innovative multimodal analytics for studying students' learning behaviors in a variety of simulation and augmented reality-based training environments. He has over 600 refereed publications, and his research is supported by funding from the Army, NASA, and NSF.*



# Proceedings of the Twelfth 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. Anne M. Sinatra is a Research Psychologist at the U.S. Army Combat Capability Development Command – Solider Center and has been a part of the Generalized Intelligent Framework for Tutoring (GIFT) team since 2012. Dr. Sinatra was the GIFTSym12 Program Chair, and is also currently the lead editor for the Design Recommendations in Intelligent Tutoring Systems book series.*

Part of the Adaptive Tutoring Series

