5-8-2018

Perceptual-cognitive Training Improves Cross-Cultural Communication

Robert Thomson
Army Cyber Institute

Michael Boyce

J.T. Folsom-Kovarik

Perceptual-cognitive Training Improves Cross-cultural Communication in a Cadet Population

J. T. Folsom-Kovarik[1], Michael W. Boyce[2], Robert H. Thomson[3]
[1] Soar Technology, Inc., [2] U. S. Army Research Laboratory, [3] United States Military Academy

INTRODUCTION

Authoring adaptive training can present challenges because instructors, unit leaders, and other non-technical users need to understand and control adaptation in order to accept and make use of a training system such as GIFT. Therefore, adaptation should be presented in a manner that parallels the way these end users think about instruction (Wray, Folsom-Kovarik, Woods, & Jones, 2015). This work enabled future improvements in authoring for adaptation by adding several constructs inside GIFT. First, *patterns* added a new construct for defining learner behaviors and analytics that can drive adaptation. Second, *misconceptions* added information to GIFT concepts in the Learner Module about reasons that individuals might be performing Below Expectation. Third, *mid-lesson reports* tested a specific type of adaptive intervention that prompts learner reflection during training, with reduced authoring via reusable prompts.

A randomized controlled trial was conducted to evaluate the training effectiveness of GIFT when driving adaptive feedback in a newly integrated tool for perceptual and cognitive skills relevant to cross-cultural communication. The combination of GIFT plus the skill training was evaluated by a population of 74 West Point Cadets. A preliminary analysis supported the value of the patterns to identify different classes of learner experience and, in future, to let non-technical personnel define what high-level behaviors and groups of observations would help GIFT respond to these. The analysis also suggested new domain-general misconceptions that might be able to inform adaptation. The evaluation showed an improvement between pre-test and post-test scores across all users. The discovery of new patterns and misconceptions highlights opportunities for instructors or unit leaders to gather evidence about how training is progressing in GIFT and, with future incorporation into the GIFT authoring suite, to quickly add new adaptive interventions that make training more effective.

Cross-cultural Communication and Perceptual-cognitive Skill Training

The proof of concept was demonstrated in a cross-cultural communication training domain. The laptop-based training consisted of four narrative scenarios that challenged learners to make decisions based on perceptual cues such as facial expressions in an environment with simulated characters. The existing training contained delayed feedback in the form of after-action review and the ability to optionally replay each scenario and try different choices. GIFT was used to add adaptive feedback to the existing training system – mid-lesson reports (see below) that were triggered by misconceptions GIFT inferred based on learner interactions. The mid-lesson reports overlaid immediate feedback onto the existing training system via the onscreen GIFT Tutor User Interface (TUI). Combining GIFT with existing training demonstrated how GIFT integration gives the potential to make a system more adaptive to learner needs by adding the tools of patterns, misconceptions, and mid-lesson reports.

We briefly describe the training content for cross-cultural communication. The scenarios and subject-matter tests in this experiment were structured around a simplified version of the Good Stranger approach to cross-cultural communication (Klein, Moon, & Hoffman, 2006). This approach is intended to work independently of a specific culture. It trains learners to perceive, understand, and work within any foreign or unfamiliar culture. The training has previously been used with success in a military setting (Hubal, van Lent, Marinier, Kawatsu, & Bechtel, 2015).

The simplified content for this experiment targeted four learning objectives:

- Initiate and engage in encounters that support the mission and build rapport
- Practice perspective-taking to make sense of encounters
- De-escalate conflicts and repair relationships
- Balance tact and tactics to achieve long-term goals in a safe manner.

Each learning objective offered opportunities during the scenarios and tests for learners to choose either Good Stranger behavior or behavior associated with misconceptions that were selected to be general and possible to express in any of the four learning objectives (see below).

Patterns

Patterns and misconceptions have the potential to add to the authoring experience in GIFT. Patterns enhance the language for expressing constraints on learner performance in the GIFT Domain Knowledge File (DKF). Patterns describe a library of reusable conditions and groups of conditions that non-technical personnel can draw on to relate domain-specific observations without requiring engineering skill to create new conditions in source code. For example, different adaptive feedback could be delivered when domain-specific conditions occur repeatedly, or close together.

A key characteristic of patterns is that they operate not on domain messages, but on conditions. Patterns let end users describe how conditions should relate to each other. So, any conditions that have already been defined by writing and compiling source code can then become part of a pattern that end users control via a future, user-friendly authoring tool. The research has previously defined an initial list of observable patterns that relate multiple performance observations together in ways instructors are likely to use (Folsom-Kovarik & Boyce, 2017). Examples include doing tasks in order or out of order, doing actions too often or too few times, and taking too long or too little time to do an action (a pattern that generalizes the GIFT slide-underdwell condition to check the timing for any pair of conditions).

This experiment led to discovery of several patterns that may be of value to instructors, which are described in the Discussion section. In advance of the experiment, the following patterns were defined:

- Hesitation: change any answer two or more times before submitting
- Hurrying: submit any answer within five seconds of a choice presentation
- Improving: exhibit Good Stranger behavior twice with no intervening mistake. This pattern enabled a positive mid-lesson report, rather than silence, when learners did well.

Misconceptions

Misconceptions enhance the GIFT learner model with additional information about estimates of mastery. Misconceptions can express not just lack of mastery but specific reasons that may underlie any incorrect or unwanted behavior that GIFT observes. By defining patterns that indicate misconceptions and adaptive feedback specific to certain misconceptions, it is possible to control in detail the feedback that GIFT adds to teaching and training.

Four misconceptions were defined for this experiment:

- Cautious: the learner is overly deferential or sacrifices a key goal
- Authoritarian: the learner is overly concerned with being respected or obeyed
- Mission-focused: the learner achieves a near-term mission at a high cost to relationships
- Rules-focused: the learner follows rules too inflexibly

The misconceptions were designed to test the reusability of the misconception idea. The same misconceptions could be expressed in all learning objectives through different learner choices or behaviors for this experiment. As a result, the GIFT pedagogical module only had to understand a single set of rules about misconceptions across scenarios. The default pedagogical module was enhanced to evaluate all misconceptions as they were inferred by conditions and patterns. For generality, the misconceptions were assigned two-dimensional values defining their importance and urgency. As a result, the enhanced pedagogical module could use domain-general rules to address the most important or urgent misconceptions first. It did not need to know about the domain-specific contents of the four misconceptions.

The reusability of these specific misconception definitions was initially limited to the cultural communication domain. In future work, it may also be possible to create misconceptions that are domain-general. Two methods might allow domain-general misconceptions. First, the domain-general misconceptions might tie to a specific instructional model and thus let GIFT infer undesirable facts about ways of learning. An example might be a misconception that it is better to avoid poor outcomes during training, when instead the specific instructional model benefits from presenting poor outcomes that challenge the learner. Second, domain-general misconceptions could be reclassified as characteristics of the learner rather than any concept. Then the GIFT learner module would update learner states and traits, rather than readiness or concept mastery. This woud give additional input to existing constructs which GIFT typically infers through surveys, such as learner grit or mastery orientation versus performance orientation.

Mid-lesson Reports

With knowledge of specific misconceptions as inferred from behavior patterns, GIFT could recommend immediate feedback that enhanced the delayed feedback already in the training. Again, reusability and generality of the approach was key. Immediate feedback was designed in the form of mid-lesson reporting. This is a form of adaptation that does not rely on information about the scenario and could be reused in any scenario during the experiment.

Examples of immediate feedback that mid-lesson reporting added to GIFT feedback (DeFalco, 2017; Goldberg, Sottilare, Brawner, & Holden, 2012) include relating good or poor performance examples to underlying reasons for performance and providing appropriate reflective prompts (Swan, 1983). Mid-lesson reports were hypothesized to improve learning outcomes by combining immediate feedback with student-directed learning and reflection. Through reflection, the learner observes the state resulting from actions and uses information from those observations to guide decisions about which actions to perform and how. To encourage reflection required including feedback on observed actions that linked the actions to target competency (Shute, 2008). This was accomplished with report messages that simply and immediately stated a possible misconception when GIFT inferred it. The report thus could be reused across scenarios, because the learners would fill in specifics from their knowledge of the action they just performed. Similar to an open learner model, the mid-lesson report would help to directly link learner choices to the computer's inference and prompt reflection on whether the inference was correct.

Three mid-lesson reports could appear for each misconception. After the third report, the wording repeated again starting from the first message. The wording of each report message was similar, with the name of the misconception differing and colored red on screen for visual differentiation. Nine positive mid-lesson reports were also available to GIFT. The mid-lesson reports let GIFT choose to challenge specific misconceptions, encourage reflection, and improve training. The design of their wording and delivery made them applicable for GIFT to deliver at any point during any scenario, for this experiment.

METHODOLOGY

The training was evaluated with a population of West Point Cadets. The authors owe a great deal of thanks to West Point faculty, staff, and Cadets for supporting and enabling this study. The population consisted of N = 74 Cadets of all years. Cadets were randomly assigned to one of two conditions, experimental or control. Each condition had 37 Cadets assigned, although one Cadet in the experimental condition either ended early or lost data due to technical failure (after the training scenarios and before the post-test and final survey). Cadets in the two conditions did not significantly differ in their performance on a subject-matter pre-test.

Study participation took 55 minutes or less per participant. The study proceeded as follows:

1. Participants read and signed an informed consent.
2. An investigator instructed the participants on the use of the training system.
3. Participants completed a demographic questionnaire and subject matter pre-test in GIFT.
4. Participants interacted with the training scenarios in order. They were allowed to review and replay each scenario if desired, but could not return to a scenario once completed.
5. Participants completed a subject matter post-test (with items identical to the pre-test for simpler balancing) and a technology acceptance survey.

The two experimental conditions differed during step 4 only. In both experimental and control conditions, the scenario content and summary feedback was the same. However, in the experimental condition, GIFT overlaid mid-lesson reporting along with the content. The mid-lesson reporting appeared in the form of tailored text messages in the GIFT TUI on the left side of the screen (presented as text only, with no character or speech). GIFT tailored the mid-lesson report messages based on choices participants made during each scenario and across scenarios. In the control condition, the TUI was always left blank.

Situational Judgment Tests

The study instrument this paper describes in detail is the pre-test and post-test. These tests were created for this study and were identical in presentation before and after the training. Each test consisted of ten written *situational judgment test* (Motowidlo & Beier, 2010) items with four to six options for each item. Participants were asked to apply their knowledge of the subject matter by stating how likely they are to try each option in the situation described. Participants could answer each option with an integer between one and six, similar to a Likert scale with excluded middle.

Each test item related to one or more learning objectives in the cross-cultural communication subject matter. Five test items were written as examples of near transfer, closely reflecting circumstances depicted in the training scenarios. Five test items were written to test far transfer. The far-transfer test items applied the same principles in new circumstances that the participants had not experienced during training. The far transfer test items were hypothesized to be valid because they were based on additional training scenarios developed by the same SMEs that were not shown to participants during the study.

Each option within a test item expressed either Good Stranger behavior or behavior associated with a specific misconception. Participants rated each option separately and were not required to choose one option or to rate one option higher than others. As a result, the pre-test and post-test yielded a large amount of data about participant skill equivalent to approximately 50 Likert items. To the extent that the options loaded on separate factors in the hypothesized five-dimensional skill application model, each dimension of that model (misconception or correct behavior) was represented on the test by between seven and ten Likert items. Seven Likert items are suggested to adequately measure most constructs (Willits, Theodori, & Luloff, 2016).

An example test item was Question 4, which produced an interestingly mixed outcome. The outcome is discussed in detail in the Results section. The text of this example test item was as follows:

> 4. A group of local contractors are compromising security to save time. You talked to them once about it but they have not changed their ways. (**LO = balance tact and tactics / Near transfer)**
> a. ______ threaten to accuse them of helping the enemy **(Rules-focused)**
> b. ______ increase patrols to backfill the compromised security **(Mission-focused)**
> c. ______ fake an attack demonstrating how compromised the security is **(Authoritarian)**
> d. ______ wait and follow up if something happens **(Cautious)**
> e. ______ call higher command **(Good Stranger)**

Additional measures captured included demographic information, cognitive load after each test and training scenario, and a final questionnaire about the training system as a whole. This paper covers initial analysis of the situational judgment tests only, while a more comprehensive analysis is still ongoing. The additional analysis will include detectable differences based on demographics if any, self-efficacy, user acceptance of the technology and training, relationship between patterns and ability to predict post-test performance, inferred misconceptions in relationship to test performance, choices made and content seen in relationship to differences in learning, and interventions presented in relationship to learning.

RESULTS AND DISCUSSION

The study found the following initial results. First, the system supported learning. Second, patterns differentiated learners. Third, inferred misconceptions aligned with ground truth as determined by the pretest and post-test. Fourth, the research suggested improvements to make this approach more general and more effective.

Training Effectiveness

Pre-test and post-test responses on the subject matter test were compared and evaluated for evidence of improvement. Participants learned and improved significantly on three of the five near-transfer test items. One test item was subject to a severe ceiling effect. The remaining test item, question 4 above, was chosen for an in-depth analysis of patterns in learner experience as described below. Figure 1 shows the outcomes of the near-transfer test items.

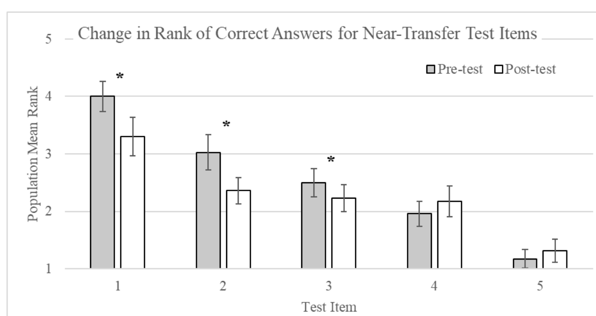

Figure 1: After training, the correct answer was ranked higher (closer to one) in 3 of 5 items.

Learning outcomes in Figure 1 were measured by the differences in scores earned on pre-test and post-test administrations of a subject matter test. The initial analysis assigns scores to test answers based on the *rank* of the Good Stranger choice as compared to other choices within the same test item. Ratings in a form such as Likert data and the answers to these test options are not scalar and therefore cannot be averaged or subjected to t-tests (Jamieson, 2004). They also cannot be compared across participants. Therefore, comparing answer rank is a more valid way to make test scores comparable and subject them to statistical tests for significance. In a test item with five options, each participant might rank the Good Stranger option first (more likely than any other) through fifth (less likely than any other). As a result, the possible ranks are comparable as real numbers between 1 and 5 inclusive, with lower numbers indicating a better score. A significant decrease in the rank score between pre-test and post-test indicates that learning occurred.

Wilcoxon rank tests were applied to the test items in Figure 1. The score improvement between pre-test and post-test was statistically significant for three of the five items (marked with asterisks). The error bars in Figure 2 indicate 95% confidence intervals. In no case do the error bars contain the extreme values (5 and 1). However, test item 5 suffered from a ceiling effect. Most participants (93%) were already rating the correct answer first in the pre-test, before any training. Fortunately, none of the other nine test items placed a majority of learners at ceiling on the pre-test.

Unlike the near-transfer questions, with the five far-transfer questions learners did not show significant improvement from pre-test to post-test. The mid-lesson reporting intervention was expected to support far transfer by making the underlying structure of the material explicit (defining the Good Stranger and the specific expected misconceptions). Further analysis is needed to determine what portion of participants received or viewed which mid-lesson reporting interventions. It is hypothesized that even in the experimental condition, participants may not have received all or enough mid-lesson reports to support effective far transfer. This can be determined with further analysis by correlating test outcomes with the specific interventions that were delivered.

Finally, test item 4 provided an interesting result which led us to examine that item further. Test item 4 may have revealed an instance of ineffective training because learners did worse on the post-test more often than they did better. This trend did not reach statistical significance but it did maintain that direction even after excluding the large number of learners (35%) who were already at ceiling on the pre-test.

A benefit of the patterns being added to GIFT is that instructors or researchers can use them to relate observations into meaningful groups and draw inferences from them. An example is finding and dealing with ineffective training after a system has been published.

Using Patterns to Detect Ineffective Training

Training can be ineffective, even if initially validated, for a number of reasons. Real-world reasons might include (1) changes in learning context that make the training less impactful or (2) changes in tactics, techniques, and procedures that make the training obsolete. Here we discuss an application of patterns to analyze learner experience in Scenario 1, which is directly aligned with the pre-test and post-test item 4.

All training scenarios used in the experiment were created by Army subject matter experts and were previously used in training experiments with Soldiers. However, an unexpected difference in training arose because of these scenarios' presentation outside the context of other training material. Specifically, Scenario 1 offers no action choices that do not result in some negative message. Either the learner's action will compromise unit safety or reduce trust with the local counterparts. The doctrinal answer is choosing to sacrifice trust to maintain security – call higher command to intervene in this case. The learner should "balance tact and tactics," and should choose to accept sacrificing tact under the circumstances.

The design of the first scenario is intended to introduce the idea that in a situational judgement test there may be a tradeoff of competing values. The choice of where to strike the balance measures which value is a higher priority, and there may be no way to simultaneously achieve all good outcomes. This idea was alluded to in the test instructions which stated: "Choices you make in this study may include limited options where there is no wholly right answer. Try to choose the best available option." However, the fact that no answer would avoid all negative outcomes was not made explicit during Scenario 1. As a result, many participants may have drawn an incorrect conclusion when they received negative feedback during the scenario. They assumed that if their first choice resulted in negative feedback, then their second guess must have been the correct one. If learners did not replay the scenario and attempt another choice, then their assumption was not challenged and created a misconception that was reflected in the post-test.

GIFT can detect patterns of behavior that suggest how different learners confronted this scenario. Examples of patterns that appeared in a post hoc analysis included completing the scenario and simply moving on without questioning. However, a relatively high percentage of learners instead changed their answer before submitting it, paged backward to review content both before and after submitting, read more of the after-action review content, and replayed the entire scenario. These patterns suggest different ways of confronting the challenging content that can be detected with patterns that pick out specific relations between multiple domain messages.

GIFT can also help to address training that becomes ineffective and help instructors make it effective again using the ability to easily author tutoring overlays in response to domain patterns. With patterns, an instructor or other non-technical author could author a new pattern that describes reading this scenario without replaying or without sufficient replaying. The pattern could trigger an overlay message that makes the tradeoff more explicit to learners. The overlay might help explain a subjective decision or might simply encourage trying other options to see what happens. In this way, a need for change in the existing content is addressed by the future pattern and overlay authoring in GIFT.

Table 1: Example paths through Scenario 1 can be expressed with patterns of order, timing, and repetition.

Example	Path
1	Cautious option, replay, mission-focused option, replay, Good Stranger option, stop
2	Cautious option, review, replay, same cautious option, review, replay, same cautious option
3	Mission-focused option, review, review, stop
4	Mission-focused option, stop

Table 1 describes a sampling of learner experiences during the choice point of Scenario 1. The number of unique paths through the single choice point, out of 74 participants, was 57. This is higher than anticipated because there was more diversity than expected in how learners engaged with the training system. New patterns should be added and tested to group these paths into meaningful categories. Table 1 suggests some possible patterns that appeared in the data and could be added to those defined in the introduction.

Example 1 describes a learner who is almost ideal in making the most of training. Whether purposefully exploring all options or simply persisting in order to get a better result, this learner plays through the entire first scenario three times, trying three different options. On the post-test, this learner was not confused by tradeoffs in the scenario about the correct answer, and ranked the correct answer highest.

Example 2 seems to describe a learner in a state of non-productive frustration or wheel spinning (Beck & Gong, 2013). When faced with negative feedback, this learner replays the scenario but tries the same option again. As if to amplify the frustration with the available options and the unwillingness to explore another path, the learner tries the same sequence a third time before quitting. Unlike the participants who tried three different options, this learner appears to need a hint to try something new.

Example 3 is the most common pattern observed in the first scenario. The learner attempts an answer, receives negative feedback, and then takes some action to learn more – but not effective action. The act of reviewing the material in this example may indicate an attempt to learn from the mistake, but the learning is not completely effective if no alternative is tried. Subject to further analysis, this pattern may be associated with worse performance on the post-test because the learner may form a belief about what the right choice was, but does not test it.

Example 4 seems quite abrupt in ending the scenario. Further analysis will determine if this pattern also includes shorter time spent reading. The failure to take advantage of the after-action review may indicate disengagement. Perhaps a more extreme example is another learner who changed the answer several times before submitting it, suggesting uncertainty, and even so ended the scenario immediately.

Since there were more paths through a single choice point than expected, and since they seem to be grouped into patterns such as those cited and others, future research should be conducted to create further patterns and test how they help GIFT infer facts about learners.

Inferring Misconceptions and Learner Characteristics

GIFT inferred misconceptions with an updated learner model structure that matched each concept with multiple misconceptions that list specific reasons a learner might be expected to be below expectations on that concept. Standard domain conditions were created to translate performance in the training scenarios to evidence of misconceptions. Presence of misconceptions enabled the pedagogical module to present targeted mid-lesson review interventions.

The accuracy of misconceptions in the learner model was measured by comparing misconception estimates against pre-test and post-test scores. In this initial report, only Scenario 1 and the matched test item 4 were analyzed. A full analysis of all scenarios and test items will be reported separately and will be able to include findings of change over time and statistical significance in the results. Note that experimental and control conditions did not differ in the method of inferring misconceptions, so the analysis includes the full population of learners.

Logs for the experiment were reviewed to determine the first misconception, if any, that GIFT detected for each learner during Scenario 1. This misconception was compared against the learner's pre-test answer to test item 4 only. If the option corresponding to this misconception was the highest ranked option on the pre-test, then GIFT was considered to have correctly identified a misconception the learner did have in mind before starting Scenario 1. The overall accuracy of the first misconception GIFT detected was 52.9%. This accuracy is higher than random chance for a five-way classification task.

Next, the experiment logs were again examined to determine the inferred misconception that GIFT estimated for each learner at the end of Scenario 1. The inferred value was compared to the learner's post-test onswer for test item 4 only. If the learner's answer gave top rank to an option reflecting the same misconception as GIFT predicted, then GIFT was considered to have correctly predicted the learner's state after Scenario 1. The overall accuracy on this test was 56.5%, which is again higher than random chance for a five-way classification task.

There were several limitations in this analysis which should be addressed by further analysis of the same data. First, only one pre-test item and one post-test item were considered. A method should be created to combine the results of all test items for each learner, in order to create a better picture of ground truth about their misconceptions. Second, the analysis does not account for learning that takes place after Scenario 1. Presumably GIFT had formed different, possibly improved estimates of learner misconceptions by the end of all the training scenarios, which should be collected and compared to find change over time. Third, the analysis only considered misconceptions detected during Scenario 1 and Scenario 1 actually did not contain an opportunity for the learners to display a rules-focused misconception. As a result, prediction accuracy was zero for all learners who really did have a rules-focused misconception in mind. Therefore a full-scale analysis may find increased accuracy by removing this handicap.

In summary, GIFT could detect misconceptions related to a specific concept. As was discovered in analysis after the experiment, it may also be possible to detect learner characteristics. Some of these characteristics already appear in the default GIFT learner model. They represent an opportunity for GIFT to update the modeling of these characteristics. These include (1) persistence or grit, (2) performance orientation versus mastery orientation, and (3) external locus of control versus internal locus of control. In addition, some characteristics suggested by the data are new and could be added to an enhanced learner model. These include (1) spinning wheels, (2) frustration or disengagement, and (3) curiosity or willingness to explore. Further research is needed to confirm that these possible traits and states are actually present in learners and are detectable by behavior patterns as they seem to be. If they are, then GIFT could have a new capability to detect domain-general "misconceptions" or facts about how people learn and use the facts to change an adaptive training experience.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

In conclusion, the present experiment demonstrated an integration of research into GIFT authoring with an actual training system and typical users. The research capabilities demonstrated were patterns that help tell the difference between individual learners based on fine-grained behavior, misconceptions that make the GIFT learner model more precise about why any learner might be below expectation, and mid-lesson reports that give a reusable tool for addressing misconceptions with reflective prompts.

The research suggests several directions for future research. First, additional research is needed to *close the loop* on mid-lesson reporting and improve the interventions to make them more effective. With future work to incorporate patterns and misconceptions into the GIFT authoring suite, it would then be possible for an instructor to define mid-lesson reports in reponse immediately on recognizing that training is not proceeding as expected. This quick response to fix ineffective training is made possible by authoring patterns and misconceptions that let GIFT capture and interpret observations in a manner similar to how instructors and end users think about training.

In addition, valuable future research could build on the success of patterns and misconceptions here in order to create an instructor-facing learning analytics capability. Analytics are a burgeoning field in adaptive training and can use the fine-grained data in GIFT to give insight into how learners are using training and what training is more or less effective. Incorporation into GIFT would also help provide instructors and unit leaders with tools to focus their training effort in real time.

Finally, in the immediate term the research will be extended to additional training domains in order to ensure the generality of patterns and misconceptions for describing training in ways that instructors want to understand. A candidate domain might be a VBS room-clearing scenario which is already created and has collected some data pertaining to learner movement in a simulated space. This kind of extension will provide an excellent opportunity to show how patterns and misconceptions let human instructors or SMEs intuitively label complex behavior in a manner that is very challenging to machine learning algorithms.

With these additional advances, the present research will add to the end-user authorability of increasingly sophisticated and effective adaptive training in GIFT.

REFERENCES

Beck, J.E., & Gong, Y. (2013). *Wheel-spinning: Students who fail to master a skill.* Paper presented at the International Conference on Artificial Intelligence in Education.

DeFalco, J.A. (2017). *Examining motivational feedback for sensor-free detected frustration within game-based learning.* Columbia University.

Folsom-Kovarik, J.T., & Boyce, M.W. (2017). *Developing a Pattern Recognition Structure to Tailor Mid-Lesson Feedback*. Paper presented at the 2017 GIFT Symposium, Orlando, FL.

Goldberg, B., Sottilare, R., Brawner, K., & Holden, H. (2012). *Adaptive game-based tutoring: Mechanisms for real-time feedback and adaptation.* Paper presented at the Proceedings of the I3M Conference on International Defense & Homeland Security Simulation Workshop.

Hubal, R., van Lent, M., Marinier, B., Kawatsu, C., & Bechtel, B. (2015). *Enhancing Good Stranger Skills: A Method and Study*. Paper presented at the Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), Orlando, FL.

Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical education, 38*(12), 1217-1218.

Klein, G., Moon, B., & Hoffman, R.F. (2006). Making sense of sensemaking I: Alternative perspectives. *IEEE Intelligent Systems, 21*(4), 70-73.

Motowidlo, S.J., & Beier, M.E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95*(2), 321.

Shute, V.J. (2008). Focus on formative feedback. *Review of educational research, 78*(1), 153-189.

Swan, M. (1983). *Teaching decimal place value: A comparative study of "conflict" and "positive only" approaches*: Nottingham University Shell Centre for Mathematical Education.

Willits, F.K., Theodori, G.L., & Luloff, A. (2016). Another Look at Likert Scales. *Journal of Rural Social Sciences, 31*(3), 126.

Wray, R.E., Folsom-Kovarik, J.T., Woods, A., & Jones, R.M. (2015). Motivating narrative representation for training cross-cultural interaction. *Procedia Manufacturing, 3*, 4121-4128.

ABOUT THE AUTHORS

J.T. Folsom-Kovarik, Ph.D. *is the lead scientist at Soar Technology, Inc. for adaptation and assessment within intelligent training. He earned a Ph.D. in computer science from the University of Central Florida in 2012. His research combines modern data science and machine learning approaches with SoarTech's long experience in modeling expert knowledge and human experience. When expert knowledge guides machine learning and data analytics algorithms, they become more applicable and useful in real-world training settings. The combination of approaches can remain feasible when available data is small, concepts evolve over time, or nontechnical users need to control the training.*

Michael W. Boyce, Ph.D. *is a research psychologist with ARL's adaptive training research program. For the past 3 years his emphasis has been in using technologies like GIFT to accurately assess learner knowledge and performance. Located at the United States Military Academy at West Point, his goal is to better inform the research progress of GIFT through interactions with a military student population. He received his Ph.D. in Applied/Experimental Human Factors Psychology from the University of Central Florida in 2014.*

Robert Thomson, Ph.D. *serves as the Cyber and Cognitive Science Fellow at the Army Cyber Institute and is an Assistant Professor in the Department of Behavioral Sciences and Leadership at the United States Military Academy. Dr. Thomson has over 7 years of post-graduate experience and over 30 invited and refereed academic publications in the domains of computational modeling, intelligence analysis, cybersecurity, and artificial intelligence. He has received funding from IARPA, DARPA, ONR, and ARL. He is also a faculty representative for the Men's Basketball Program at West Point.*