Learning, Adaptive Support, Student Traits, and Engagement in Scenario-Based Learning

Mark G. Core, Kallirroi Georgila, Benjamin D. Nye, Daniel Auerbach, Zhi Fei Liu, Richard DiNinni
University of Southern California, Institute for Creative Technologies
Playa Vista, CA
core,kgeorgila,nye,auerbach@ict.usc.edu, zhifeili@usc.edu, dininni@ict.usc.edu

ABSTRACT

Scenario-based training systems pose an especially difficult challenge for an intelligent tutoring system (ITS). In addition to the basic problems of deciding when to intervene and what guidance to provide, the ITS must decide whether to give guidance directly (e.g., a hint message), indirectly through positive/negative results in the scenario, or to delay guidance until a post-scenario review session. There are a number of factors that an adaptive ITS should consider and we use self-report survey instruments to investigate the relationship between traits, learning strategies, expectations, learner behaviors derived from log files, post-use perceptions of the system, and pre-test and post-test results. We use the ELITE Lite Counseling training system as a testbed for our experiments. This system uses virtual role players to allow learners to practice leadership counseling skills, and is in use at the United States Military Academy (USMA). This paper analyzes two data sets. We collected data from local university students, a non-military population of roughly the same age as USMA Cadets using the system. For these local participants, we could administer surveys and pre-tests and post-tests, and collect log files recording clicks made while using ELITE Lite. The second data set comes from USMA itself but is limited to log files. In both populations, the ITS's hints are effective at boosting scenario performance, and for the university students, the overall experience promoted learning, and survey results suggest that higher levels of organization in study habits may lead to greater learning with ELITE Lite. For the USMA Cadets, ELITE Lite is part of their Military Leadership course rather than an experiment, which could explain why we found higher scenario performance on average than the non-military population, and more use of the post-scenario review feature.

ABOUT THE AUTHORS

Mark G. Core, Ph.D. is a Research Scientist at USC ICT. He received his Ph.D. from the University of Rochester in 2000, and was a Research Fellow at the University of Edinburgh until joining USC ICT in 2004. Mark specializes in artificial intelligence in education using virtual humans as role players for interpersonal skills training (e.g., negotiation, cultural awareness and leadership). His research focuses on how best to support learners using automated guidance, and support authors creating realistic and engaging virtual humans that promote learning.

Kallirroi Georgila, Ph.D. is a Research Scientist at USC ICT and Research Faculty at USC's Computer Science Department. Before joining USC ICT in 2009 she was a Research Scientist at the Educational Testing Service in Princeton, and before that a Research Fellow at the University of Edinburgh. Kallirroi's research interests include all aspects of human-machine dialogue processing with a particular focus on machine learning for dialogue management.

Benjamin D. Nye, Ph.D. is the Director of Learning Sciences at USC ICT. He received his Ph.D. in Systems Engineering from the University of Pennsylvania in 2011 and previously served as Research Faculty at the University of Memphis. Ben's major research interest is to identify best practices in advanced learning technology, particularly for frontiers such as distributed learning technologies (e.g., cloud-based, device-agnostic) and socially-situated learning (e.g., face-to-face mobile use). Ben's research tries to remove barriers to development and adoption of intelligent tutoring systems so that they can reach larger numbers of learners, which has traditionally been a major roadblock for these highly-effective interventions.

Daniel Auerbach is a Research Programmer at USC ICT. He received a B.A. in Linguistics and Computer Science from Cornell University.

Zhi Fei (Emily) Liu was a student research worker at USC ICT and currently works for Microsoft. She received a B.S. in Computer Science from USC.

Richard DiNinni is a Project Director at USC ICT and has played a role in the development and transition of a diverse set of training programs over the past 15 years. Most recently he has been leading a partnership with the United States Military Academy and serving as a visiting researcher. He received a M.S. in Communication from Syracuse University.

Learning, Adaptive Support, Student Traits, and Engagement in Scenario-Based Learning

Mark G. Core, Kallirroi Georgila, Benjamin D. Nye, Daniel Auerbach, Zhi Fei Liu, Richard DiNinni
University of Southern California, Institute for Creative Technologies
Playa Vista, CA
core,kgeorgila,nye,auerbach@ict.usc.edu, zhifeili@usc.edu, dininni@ict.usc.edu

INTRODUCTION

Traditionally, designers of training systems have focused on task performance (i.e., does practice in the system lead to improved performance) with little concern for learner emotions. However, recent thinking about best practices for instruction acknowledges a strong link between cognition and emotion, with D'Mello and Graesser (2012b) presenting evidence for a model of learning comprised of cognitive-affective states (engagement, confusion, frustration, and disengagement). Of these, while engagement and confusion can be productive, disengagement is known to impact learning negatively. Baker, D'Mello, Rodrigo, and Graesser (2010) reported that disengagement is associated with gaming the system (e.g., random guessing, hint abuse). In addition, disengagement can affect attitudes toward the subject matter and even school itself, as measured by dropout rates (Christenson, Reschly, & Wylie, 2012). These factors interact with traditional cognitive factors as well, such as cognitive load and feedback signals (Graesser, 2009). However, these effects may be complicated by the fact that different student traits (e.g., level of organization) and educational contexts (e.g., high stakes vs. low stakes) are likely to moderate the expression of academic emotions and their impact on learning.

These issues are particularly relevant to rich learning environments where learners encounter realistic problems, such as in a scenario-based intelligent tutoring system (ITS). Adaptively-scaffolded scenarios can provide engagement opportunities and realistic practice in domains ranging from traditional classroom education (Rowe, Shores, Mott, & Lester, 2011), combat training (van Lent, Fisher, & Mancuso, 2004) and interpersonal skills training (Kim et al., 2009), but little is known about how different types of users benefit from common elements such as dynamic hints and reflection as a result of after-scenario (or after-action) reviews (AARs). Scenario-based ITSs have a large design space, where guidance can be provided in multiple ways. Such guidance might be delivered as consequences in the scenario (e.g., through a simulation) or through other mechanisms (e.g., a tutor or coach agent). There can be variations in the timing of support (e.g., proactive, reactive, or delayed) or the initiative for feedback (e.g., user-requested or system-generated).

In this paper, we study the interactions between learning, adaptive support, student traits, and engagement in the context of a scenario-based ITS for interpersonal skills training. This system, called ELITE (Emergent Leader Immersive Training Environment) Lite Counseling, allows learners to practice leadership counseling with virtual role players, and is in use at the United States Military Academy (USMA). To explore these interactions in depth, we conducted an experiment with a comparable non-military population of university students using surveys and a pre-test/post-test design. Results from this university study population were also compared with data collected in-vivo from USMA. Comparisons were made between the populations using measures derived from behavioral data stored in log files since survey and pre-test/post-test data was not available from USMA. As expected, USMA Cadets using ELITE Lite as part of their Military Leadership course showed higher performance on average than non-military participants in an experiment, though ceiling effects were still not encountered.

BACKGROUND AND RELATED WORK

Given the goal of promoting learning and productive academic emotions, and using available resources (e.g., a coach and AAR), an ITS must make hundreds of decisions per scenario about when to provide guidance, how to provide guidance (e.g., through the scenario, through the coach), and what content to provide. Although best practices for instruction exist, they are high-level and often involve trade-offs. Consider the 25 learning principles outlined by Graesser (2009). A common trade-off is keeping cognitive load manageable but providing immediate feedback. Another trade-off is breaking complex tasks into manageable parts but also making a clear link to real world examples which are often complex and messy. Understanding these trade-offs requires analyzing how a system interacts with the cognition, emotions, and behavior of students.

Prior research on the benefit of and engagement with elements of an ITS has typically focused on either physiological measurement or behavioral analysis. Physiological measurement has been particularly effective for identifying the dynamics of student emotions, e.g., D'Mello and Graesser (2012a) and for mapping out the relationship between different emotions and learning, e.g., Baker et al., (2010). Physiological and human-observable measures of engagement and emotions have also been used to attempt to influence learner emotions (e.g., facilitating regulation or positive attitudes). D'Mello and Graesser (2012a) compared a Supportive AutoTutor version that used empathetic messages against a Shake-Up AutoTutor that was informal and funny. The Gaze Tutor (D'Mello, Olney, Williams, & Hays, 2012) asks students to pay attention if its eye-tracker indicates they are looking in a region that is not relevant to the current tutoring activity. Lehman et al. (2011) studied a tutoring system designed to induce confusion in order to promote learning through the resolution of the confusion.

Conversely, behavioral analysis typically focuses on analyzing event streams for behaviors hypothesized to be productive or unproductive for learning, such as gaming the system (Baker et al., 2010), without-thinking-fastidiously (Wixon, Baker, Gobert, Ocumpaugh, & Bachmann, 2012), off-task behavior (Rowe et al., 2011), or time-on-task anomalies, such as skipping through content (Beck, 2005). Behavioral analysis has been particularly useful for identifying elements of an ITS where learners are not using the system as expected, potentially reducing learning. The work presented in this paper focuses on behavioral analysis, since it is hypothesized that different feedback mechanisms and learner characteristics may modulate learning and engagement with respect to different aspects of a scenario-based ITS.

Significant evidence exists that learners may require different types of adaptive support and feedback, either due to learner states (e.g., temporary reactions to content), longer-term student characteristics, or different advantages for different support mechanisms (e.g., immediate vs. delayed; in-task vs. coached). The interaction between temporary states (e.g., disengagement with certain aspects of the system) and longer-term student characteristics is sometimes murky. For example, Baker et al. (2010) reported that disengaged learners demonstrated poor help-seeking skills or hint abuse until the ITS gave away the answer. However, this might alternatively mean that learners who are more prone to hint abuse are also less likely to want to engage with content. This hypothesis is partly supported by the fact that gaming behavior tends to be non-normal (e.g., a small number of students do most of the gaming) and also that prevalence of certain off-task behaviors has been found to vary across cultures (Rodrigo, Baker, & Rossi, 2013). This implies potentially distinctly different needs for adaptation: 1) Delivering domain knowledge to facilitate learning, 2) Helping to regulate the dynamics of transient academic emotional states such as disengagement (e.g., like the Supportive AutoTutor; D'Mello & Graesser, 2012a), and 3) Adapting ITS activities or feedback to disincentivize persistent unproductive behavior (e.g., reacting to gaming the system; Baker et al., 2006).

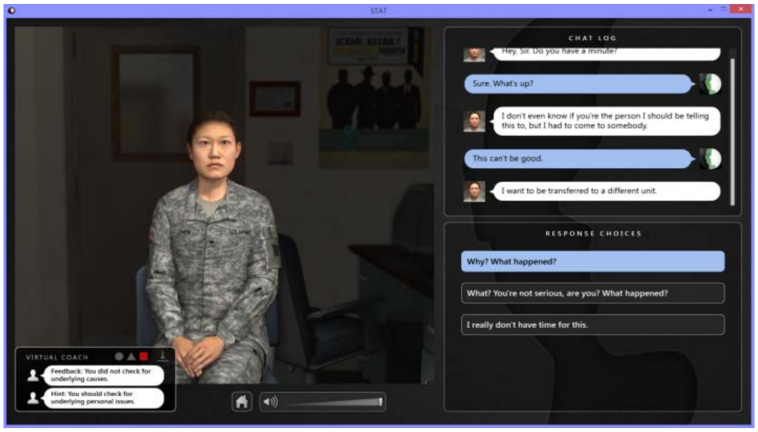
Thus, in the context of scenario-based learning, the ITS must constantly answer questions such as: When to intervene? How to intervene (e.g., through a coach, through an AAR)? What problem-solving, metacognitive and/or emotional support to provide? One possible solution is to use an empirical approach such as reinforcement learning to develop a tutoring policy to address these questions (Chi, Jordan, & VanLehn, 2014). However, such bottom-up approaches often provide little generalizable knowledge about the theoretical constructs that lead to better learning interactions. To understand these mechanisms, it may be crucial to either measure or infer traits such as help-seeking skills and expectations about ITS usefulness. A truly adaptive ITS may need to behave very differently depending on the target learner, but more research is needed to explore what interactions are important and why.

LEARNING PLATFORM COMPONENTS

We are investigating these issues in a scenario-based training system called ELITE Lite Counseling. ELITE Lite is a PC application developed from a complex multi-computer system called ELITE/INOTS (Hays et al., 2012). Both systems help officers in training develop the necessary interpersonal skills to help subordinates with personal and performance problems, and are in use at multiple military sites. The systems reinforce a set of core skills: active listening, checking for underlying causes of the problem, asking additional questions and verifying information, and responding with a course of action. On a test developed to align to ELITE/INOTS content, Hays et al. (2012)

reported significant increases in test scores after training which included INOTS, with no sign of a ceiling effect. Thus, ELITE Lite is relevant and helpful based on these measures, but there is room for improvement.

Using ELITE Lite, learners view instructional videos, and then play interactive scenarios. The scenarios allow a learner to practice skills with virtual characters (Fig. 1). Characters speak their utterances using pre-recorded audio and a transcript appears on the upper right. Learners interact with the character by selecting utterances from the menu in the lower right corner. These choices lead to different nodes of a branching graph. Scenario authors can annotate choices as positive examples of skills that are either required or optional at this point in the conversation. Negative annotations of choices include failing to apply a required skill or exhibiting a misconception. Based on these annotations, choices are classified as correct (only positive annotations), incorrect (only negative annotations), or mixed (both positive and negative annotations).



Hey, Sir. Do you have a minute?

Sure. What's up?

I don't even know if you're the person I should be telling this to, but I had to come to somebody.

This can't be good.

I want to be transferred to a different unit.

Why? What happened?

What? You're not serious, are you? What happened?

I really don't have time for this.

Feedback: You did not check for underlying causes.

Hint: You should check for underlying personal issues.

Figure 1. ELITE Lite Scenario

The character gives implicit feedback through their responses. Correct choices tend to prompt the character to provide more information and exhibit more positive emotions. There is also a coach who communicates with the learner through the chat window in the lower left corner of Fig. 1. The shapes on the title bar of the coach window light up in green, yellow or red, as a form of flag feedback for correct, mixed, and incorrect choices, respectively. Textual feedback and hints from the coach may also be presented for mixed or incorrect choices. Feedback points out either a positive or negative aspect of the last choice while hints reveal a skill associated with the correct choice of the next decision. The coach is driven by heuristics and may give both feedback and a hint or just give one to avoid repeating itself (e.g., not give feedback on active listening and give a hint to actively listen).

After each simulated conversation, learners receive an after-action review. The AAR begins with a summary screen consisting of a score, performance graphs and feedback text. Following the summary screen is a self-directed tool that enables learners to examine decisions they made and see how the available choices relate to specific skills. This tool highlights key mistakes, but learners can review any of the decisions they made.

EXPERIMENTAL DESIGN

The data for the in-depth university study analysis was collected as a formative baseline study of the system, with two randomly-assigned conditions, with the only difference being whether or not learners received textual coach guidance when they provided a mixed answer. The study design was a pre-test/post-test design, where learners completed (in this order) a pre-survey, a pre-test, the ELITE Lite intervention, a post-survey, and a post-test. Surveys were used to collect data on student characteristics, since we hypothesized that student behavior will be influenced by traits and metacognitive skills.

Items on expectations and motivation were drawn from the Modified Attitudes Toward Tutoring Agents Scale (M-ATTAS; Jackson, Graesser, & McNamara, 2009). These were supplemented by items from the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich, Smith, García, & McKeachie, 1993) that measures motivation interest, confidence, and anxiety. MSLQ items on learning strategies also queried subjects' organization skills, help-seeking tendencies, and study effort. In addition, questions were asked about the subjects' growth mindset (Dweck, 2006), i.e., the expectation that intelligence can change. These item sets were reduced from the full measures, taking the items with the highest loading for the construct (except in cases where the original survey only had one item, such as the M-ATTAS question "Computers can help me learn difficult course concepts").

After subjects interacted with the ITS, we gauged their impressions with variants of items from the Academic Emotions Questionnaire, AEQ (Pekrun, Goetz, Titz, & Perry, 2002) and the Unified Theory of Acceptance and Use of Technology, UTAUT (Venkatesh, Morris, G. Davis, & F. Davis, 2003). We used AEQ items to measure interest and negative emotions such as boredom and discouragement. We used UTAUT items to measure positive emotions (e.g., fun) and impressions of system performance (e.g., it helped me learn more quickly). We also used UTAUT items on usability (e.g., easy to use) because interface problems could influence outcomes independently of the inner workings of the ITS. UTAUT items were applied to each major component of the system (e.g., videos, character, the coach, and the AAR). Scales for each survey were placed on a 6-point Likert scale (completely disagree, mostly disagree, slightly disagree, slightly agree, mostly agree, completely agree).

For the test used as a pre-test and post-test, items were adapted from a previous study (Hays et al., 2012) of ELITE/INOTS (a classroom-level intervention) by changing military-specific elements to a workplace setting. These questions included basic knowledge questions (i.e., true-false questions) and situational judgment questions (i.e., questions about the appropriateness of different actions in four scenarios). We report three measures from these tests: percent/proportion of responses that were correct (0-1) on the knowledge questions alone, the situational judgment questions alone, and on the entire test. In addition to the tests, we also adapted questions about self-reported experience and confidence. Note that unlike our other survey items, the confidence questions used a 7-point Likert scale (e.g., 1="I am certain I cannot," 4="I am unsure if I can" and 7="I am certain I can").

Experimental Procedure for Collecting Non-Military Data

A custom-version of ELITE Lite implemented the full experimental procedure, where each participant could take up to two hours to complete the study (however, the maximum observed duration was under 1.5h). Data was collected through extensive log files of all interactions (both with the agents and the AAR), online surveys, and webcam recordings (annotation of these recordings is ongoing). We randomly assigned each participant either to the always-mixed-guidance condition (feedback and hint text given for mixed responses) or the no-mixed-guidance condition (only color-coded feedback given for mixed responses). In both conditions, guidance was given when a wrong (not mixed) answer was provided. These conditions were expected to be largely equivalent (i.e., a baseline study), with some chance of identifying a possible difference in learning gains. Participants were not aware of this manipulation and without knowledge of the system the difference would not be particularly noticeable.

Two scenarios were used for this procedure, which were chosen due to their relevance to a broad range of populations. The first scenario, *Being Heard*, involves a subordinate asking for help with a problem whose root cause is harassment (shown in Fig. 1). Participants completed Being Heard twice to potentially correct mistakes and learn from their first attempt. The second scenario, *Bearing Down*, involves a Soldier who has grabbed and threatened another Soldier. Because the coach is not active, this scenario acts as an in-game assessment of skills for participants. See Fig. 2 for the full experimental protocol.

1. Pre-survey, followed by the pre-test.
2. Introductory video on basic counseling skills.
3. Being Heard scenario first try (i.e., Being Heard 1).
4. AAR for the Being Heard scenario first try.
5. Being Heard a second time (i.e., Being Heard 2).
6. AAR for the Being Heard scenario second try.
7. Complete the Bearing Down scenario, with no color feedback or coach guidance.
8. AAR for Bearing Down.
9. Post-survey, followed by the post-test.

Figure 2. Experimental Procedure

Non-Military Participant Demographics

Data collection was conducted at a highly competitive, private United States university during the fall semester of 2015. Students were recruited using online advertisements (Facebook) and financially compensated for participating. Out of 80 participants, the data for 6 was corrupted due to software failure, resulting in a final data set including 74 students (54 male and 20 female). Of these 74 participants, 39 were in the no-mixed-guidance condition and 35 were in the always-mixed-guidance condition. Possibly due to the nature of the study (e.g., virtual agent scenarios) and online recruiting mechanism, participants tended to be from technical majors (e.g., computer science). The majority of participants (63) self-identified as “Asian/Pacific Islander” with 5 “White,” 4 “Other/did not respond,” and 2 as “Black/African American.” There were no significant differences between genders for any of the analyses, and sample sizes for ethnicity were insufficient for analysis.

USMA Data Collection

ELITE Lite is a lab exercise in the PL 300, Military Leadership, course at USMA, which all cadets must take in their third year. After receiving classroom instruction on a set of key principles for counseling subordinates, the cadets go to the West Point Simulation Center for lab sessions where they practice applying those principles using ELITE Lite, once near the beginning of the semester and again near the end of the semester. All members of the class are required to complete the Being Heard scenario and email a screenshot of the AAR Summary to their instructor. Prior to the second lab session, instructors go through a scenario in class and discuss the menu choices, provide personal experiences where appropriate, and allow the cadets to share their opinions. We received an anonymized corpus of log files from the Sim Center collected in fall 2015. Although no survey information was collected, ELITE Lite asks users to identify themselves as male or female so that the virtual role player can use gender appropriate language (e.g., “Sir” vs. “Ma’am”). In this corpus, we found 392 pairs of log data for Being Heard (i.e., same person playing at the semester beginning and end) of which 330 were male and 62 female. In the analyses below, we compare the behavioral data from these two Being Heard sessions to the two attempts at Being Heard in our local data collection.

ANALYSIS AND RESULTS

Pre-Survey: Student Characteristics

When two or more items were included for a single survey measure and taken from the same source, these were averaged prior to analyses. Average values for pre-survey measures are presented in Table 1. Note that all results were transformed such that higher values always have a positive connotation (e.g., we report lack of anxiety). Based on these results, we see that subjects reported high motivation. Although participants did not report much experience, they were highly confident as reflected in the ELITE-survey confidence items and the MSLQ question on confidence in their ability to master the material. Subjects rated their interest in the experience high, both as a study (M-ATTAS) and to learn the material (MSLQ). Average scores on “lack of anxiety (about tests)” and “effort (I put into classes)” were lower but close to “slightly agree.”

Table 1. Pre-Survey Means of Student Measures which Average N Items

Measure	N	Adapted from	Mean	95% CI
Growth Mindset	4	(Dweck, 2006)	4.10	3.86-4.35
Comfort w/ Computers	1	(Jackson et al., 2009)	5.73	5.6-5.86
Computers Can Help Learning	1	(Jackson et al., 2009)	5.27	5.10-5.44
Interest	2	(Jackson et al., 2009)	5.32	5.11-5.53
Interest	2	(Pintrich et al., 1993)	5.22	5.02-5.43
Lack of Anxiety	2	(Pintrich et al., 1993)	3.66	3.36-3.95
Effort	2	(Pintrich et al., 1993)	3.70	3.43-3.98
Help seeking	2	(Pintrich et al., 1993)	3.73	3.55-3.91
Organization	2	(Pintrich et al., 1993)	4.93	4.72-5.13
Confidence	1	(Pintrich et al., 1993)	4.66	4.41-4.92
Confidence	6	(Hays et al., 2012)	5.56	5.38-5.75
Experience	2	(Hays et al., 2012)	3.31	3.0-3.62
Video Game Frequency	1	N/A	3.71	3.27-4.16

The average scores for help seeking, growth mindset and organization were lower than interest or confidence scores, which were almost uniformly high. Unexpectedly for this population, average agreement with “I play video games frequently” was relatively low, despite agreement with “I am comfortable using computers” being very high. The average response to “Computers can help me learn difficult course concepts” was also high.

Perceptions of ITS Components

Perceptions of different ITS components were largely positive but uneven at points. Self-reported usefulness was positive for each of the ITS components and for the system as a whole. However, the coach was rated much less useful than the others with the following mean ratings: AAR (5.41), characters (5.30), system (5.36), video (5.19), and coach (4.55). Pairwise comparisons using the Wilcoxon signed rank test showed that the coach had a significantly lower average rating than the system and the other components (p<0.001). Emotional reactions to the characters and the coach (from the AEQ and UTAUT: e.g., fun, annoying, boring) found a similar difference, with average positive emotions significantly higher for the characters than for the coach (5.15 vs. 4.72, p<0.001, Wilcoxon signed rank test).

For the characters, coach and AAR, we asked participants to rate the following statement "I know how well I was doing (based on this component)." The mean responses were: characters (4.62), coach (4.73) and AAR (5.39). Pairwise comparisons using the Wilcoxon rank sum test (due to missing values the signed rank test was not used) showed that the AAR had a significantly higher average rating than the characters and the coach (p<0.001). It should be noted that the high rating of the AAR could reflect the score report and performance graph which are clear, direct performance indicators. It is interesting that the flag feedback of the coach did not lead to a similarly high rating.

Jackson et al. (2009) found a positive relationship between high agreement with "Computers can help me learn difficult course concepts" and post-tutoring impressions. This relationship was replicated in this study, where it correlated significantly (at least p<0.05) with a number of other post-tutoring impressions such as positive emotions about the characters (r=0.33), the coach (r=0.22), and the system overall (r=0.38). The perceived usefulness of the video content was also significantly correlated with this item (r=0.38).

Behavioral Engagement and Timing

Engagement, as measured by time on task, was greater for the first attempt at the Being Heard scenario than the second attempt. Average response times for Being Heard 1, Being Heard 2, and Bearing Down were 13s, 8.28s, and 14.07s respectively (all differences pairwise significant at p<0.001, Wilcoxon signed rank test), so 36% less time was spent in making decisions on the second attempt of Being Heard.

AAR use was low overall and was lower for each subsequent scenario. The mean AAR durations for the two Being Heard scenarios were significantly different (23.54s vs. 19.23s; Wilcoxon signed rank test, p<0.001). There was a drop in AAR duration for the Bearing Down scenario, which was barely used at all (13.16s; Wilcoxon signed rank test, p<0.001 between Being Heard 1 and Bearing Down and p<0.05 between Being Heard 2 and Bearing Down). The total number of clicks on the interactive AAR was very low, averaging 0.12 clicks per scenario, which essentially indicates no use. So, while learners self-reported strongly positive views about the AAR, these may be merely reactions to the summary scores rather than productive reflection on their specific mistakes.

Moreover, all surveyed student characteristics were almost entirely uncorrelated with use of the AAR. Among student characteristics, a Dweck growth mindset correlated with less time on the final AAR (r=-0.24; p<0.05). More time on the interactive AAR was correlated with higher confidence gains (r=0.36; p<0.01), which appears to be primarily influenced by additional time on the second Being Heard attempt (r=0.41; p<0.001). However, such confidence does not necessarily indicate better learning: the relationship between learning gains and additional time on the interactive AAR was not significant (which, given the lack of clicks to explore it, could be expected).

Time-based indicators were also associated with scenario performance. Response time for correct responses was negatively correlated for Being Heard 1 scores (r=-0.27; p<0.05) and Bearing Down scores (r=-0.26; p<0.05), but not for Being Heard 2 (the second attempt for Being Heard) scores. Such delayed responses may represent an indicator of confusion or uncertain knowledge.

For the USMA data, the initial average response time for Being Heard 1 was similar, 13.59s, but fell less drastically for Being Heard 2, 11.93s. The differences between the two USMA averages are significant (p<0.001, Wilcoxon signed rank test). USMA learners spent more time on average in the AAR than the non-military population, and the time increased from the first practice to the second. For the first practice, the average time was 87.69s (average #

clicks was 2.1) and for the second practice, 245.96s and 1.99 clicks. The differences between these averages is significant (p<0.001 for AAR time and p<0.05 for clicks, Wilcoxon signed rank test). Time spent to select a correct response had a negative correlation with score (r=-0.10, p<0.05) for the second Being Heard practice but no relationship was found between these variables for the first Being Heard practice.

Performance, Learning and Confidence

As shown in Table 2, non-military participants also demonstrated learning gains overall, for shallower knowledge questions (i.e., true-false classification), and for deeper situational judgment test questions (SJT). Situational judgment, which had the highest alignment to the scenario-based ITS instruction and which was initially more difficult, showed a greater gain than shallow knowledge. Although there were significant gains between pre-tests and post-tests, in all cases, average scores on both types of test items remained less than 0.70 after the intervention (which lasted approximately 30m to 1h). There was no significant difference in the learning gains between the two randomized conditions that varied coach behavior in response to subjects selecting mixed choices.

Table 2. Test Scores and Learning Gains

	Pre-test	Post-test	Learning Gain
Combined	0.57	0.65	0.08***
Knowledge (Shallow)	0.60	0.64	0.04***
SJT (Deep)	0.55	0.66	0.11***

*** Pre- and post-test significantly different at p<0.001, Wilcoxon signed rank test

Regression tests showed that the learning gain for shallower knowledge questions is dependent on the pre-test knowledge score (p<0.001) and pre-survey anxiety (p<0.05). The learning gain for deeper situational judgment test questions (SJT) is dependent on the pre-test SJT score (p<0.01) and pre-survey organization (p<0.01). Similarly, the combined (knowledge and SJT) learning gain is dependent on the pre-test combined score (p<0.01) and pre-survey organization (p<0.01). With that said, 10-fold cross-validation testing indicated that these models were only weak predictors of learning gains ($R^2<0.05$).

Performance on the scenario-based ITS was higher for the first scenario type (Being Heard), where coach support was available. Perfect performance in Being Heard (first scenario) and Bearing Down (final scenario) would produce scores of 19 and 20 (i.e., the minimal perfect dialogue path for each scenario) where correct choices earn 1 point and mixed choices earn 0.5 points. Pairwise differences between all scenarios were significant (p<0.001, Wilcoxon signed rank test). Being Heard was attempted twice with coach support. While average performance increased across these sessions (15.2 to 16.4), it did not come close to perfect, despite the ability to use the AAR to review every incorrect decision. Bearing Down had an average score of 12.4, which is lower than Being Heard. Performance on Bearing Down correlated moderately with both SJT and combined post-test performance (further details are given in the next section). Conversely, performance on Being Heard did not correlate with any component of post-test scores.

Despite the fact that participants did not perform particularly effectively on either the tests or the scenarios, their confidence was high. As noted in Table 1, confidence in performance started high (5.56 out of 7). Confidence also increased significantly on the post-survey (6.0; p<0.001; Wilcoxon signed rank test). While pre-test ELITE confidence correlated with combined pre-test performance (r=0.24; p<0.05), this relationship did not hold for the post-test and post-survey. Instead, post-survey confidence remained highly correlated with pre-survey confidence (r=0.78; p<0.001), rather than changing due to accurate self-assessment of learning and performance.

For the USMA data, average scores were higher: 16.5 for Being Heard 1, and 17.34 for Being Heard 2. There was a significant difference between the two USMA practice sessions (p<0.001, Wilcoxon signed rank test). There was also a significant gender difference with females scoring higher on average across both Being Heard practice sessions (Being Heard 1, male=16.4, female=16.8, p<0.01, Wilcoxon rank sum test; Being Heard 2, male=17.25, female=17.8, p<0.05, Wilcoxon rank sum test).

The Impact of Adaptive Support

The experimental manipulation resulted in more hints for the always-mixed-guidance condition compared to the no-mixed-guidance condition (p<0.001, Wilcoxon rank sum test) for both scenarios where support was given (Being

Heard 1 and Being Heard 2). The version of the system used at USMA is between these two extremes and gives guidance for mixed choices if the current performance is below a fixed threshold. Table 3 shows the mean of the number of hints and feedback for each Being Heard scenario.

The magnitude of the differences between the no-mixed-guidance and always-mixed-guidance conditions (close to 3 times as much support) indicates that learners were much more likely to select mixed choices than completely wrong choices. This difference was relevant to learning gains between conditions, with higher SJT gains correlated with selecting mixed choices in the no-mixed-guidance condition (r=0.33; p<0.05) but not in the always-mixed-guidance condition. This effect may be associated with the negative correlation between the probability of selecting mixed responses and lower pre-test scores (r=-0.24; p<0.05).

Table 3. ITS Coach Support by Scenario and Condition

Hints	**no-mixed-guidance**	**always-mixed-guidance**	**USMA**
Being Heard 1	0.69	2.77	2.31
Being Heard 2	0.56	1.86	1.99
Total	1.26	4.63	4.30
Feedback	**no-mixed-guidance**	**always-mixed-guidance**	**USMA**
Being Heard 1	0.97	2.8	0.78
Being Heard 2	0.51	1.86	0.53
Total	1.49	4.66	1.30
Total Support	2.74	9.29	5.60

Scenario performance was significantly higher for always-mixed-guidance in both Being Heard scenarios (p<0.01, Wilcoxon rank sum test) but not for Bearing Down, which provided no feedback or hints for either condition. Specifically, the additional guidance appeared to impact the probability that the learner will provide a correct or incorrect response (but not a mixed response). Table 4 shows the average probabilities that the learner will provide a correct, mixed, or incorrect response given a hint. For comparison, we can also see the overall probabilities (given or not given a hint). Hints increased the probability of correct responses. Furthermore, for Being Heard 1, there was a higher probability of correct responses in the always-mixed-guidance condition compared to the no-mixed-guidance condition (p<0.05, Wilcoxon rank sum test). There was a similar situation for Being Heard 2 but it did not reach significance. Likewise, the probability of incorrect responses was significantly lower for the always-mixed-guidance condition compared to the no-mixed-guidance condition for both Being Heard 1 (p<0.001, Wilcoxon rank sum test) and Being Heard 2 (again, not significant). There was no significant difference with respect to the probability of mixed responses between the two conditions. On average, responses took longer in the no-mixed-guidance condition, but this was significant only for Being Heard 2 (p<0.05, Wilcoxon rank sum test). Despite these differences on immediate performance, condition membership did not significantly impact combined learning gains.

Table 4. Impact of Hints on Response Quality

	Non-military			**USMA**		
Given a Hint	**Prob Correct**	**Prob Mixed**	**Prob Incorrect**	**Prob Correct**	**Prob Mixed**	**Prob Incorrect**
Being Heard 1	0.86	0.07	0.07	0.91	0.05	0.04
Being Heard 2	0.89	0.05	0.05	0.92	0.05	0.03
Overall	**Prob Correct**	**Prob Mixed**	**Prob Incorrect**	**Prob Correct**	**Prob Mixed**	**Prob Incorrect**
Being Heard 1	0.79	0.16	0.05	0.85	0.11	0.04
Being Heard 2	0.86	0.10	0.04	0.88	0.10	0.03

As noted above, performance on Bearing Down was significantly worse than for Being Heard. This may be due to the absence of coach and/or increased difficulty of the scenario. Both scenarios were designed to be equally difficult but in practice some mixed and incorrect responses may be better “distractors” than others. However, there is some evidence that the lack of coach support may explain part of this difference: a students’ probability of correct answers on Bearing Down correlated moderately with both SJT (r=0.23; p<0.05) and combined post-test performance (r=0.29; p<0.05). Conversely, Being Heard scores did not correlate with any component of post-test scores. This indicates that without coach support, the scenario may be more difficult and function more like an assessment.

For the USMA data, the coach is always active and offering guidance after mixed choices depending on the user’s performance. In Table 4, we see that overall the cadets are performing higher (i.e., higher probability of correct choices) than the non-military population, and that in both populations hints cause performance to improve. For the USMA data, there is a gender difference. Across both Being Heard practice sessions, the mean probability of a correct choice is significantly different for males, 0.86, and females, 0.89 (p<0.001, Wilcoxon rank sum test). For quantity of support, there are gender differences in the USMA data. For feedback on Being Heard 1 and 2, males had an average of 1.41 and females, 0.74. For hints on Being Heard 1 and 2, males had an average of 4.44 and females, 3.58. Both differences are significant (p<0.001 and p<0.05 respectively, Wilcoxon rank sum test).

CONCLUSIONS AND FUTURE DIRECTIONS

Overall, non-military participants demonstrated significant learning: a 14% improvement on combined scores and 20% improvement on situational judgments. In both populations, improvement in scenario performance was also observed. The average score on Being Heard increased in the non-military population from 15.2 to 16.4, and in the USMA population from 16.5 to 17.34. The ability to teach such interpersonal skills using an ITS offers opportunities to help learners practice difficult scenarios without risking harm to real-life peers or subordinates. This analysis indicates that ELITE Lite measurably improves key basic counseling skills. However, the interactions between learning, adaptive support, student traits, and behavioral engagement were nuanced in this analysis. Behavioral indicators of post-test performance mostly functioned as de-facto intermediate assessments. These included performance on Bearing Down (the scenario with no feedback) which predicted higher post-test scores, as well as the probability of selecting mixed responses in the no-mixed-guidance condition, which correlated with both pre-test and post-test scores. Constructs derived from student self-report (e.g., growth mindset, MSLQ, perceptions of the system components, etc.) tended to correlate among each other, but few correlated significantly with either learning gains or with behavioral indicators of engagement (e.g., productive use of the AAR).

When analyzing both student characteristics and behavioral engagement against other factors, one confound was that the non-military population was relatively homogeneous with respect to self-report and engagement. Behavioral engagement of students in the system was quite similar, in terms of their time to complete the elements of the study and in their limited use of the AAR. Likewise, for many self-report items, participants literally used only half of the scale (i.e., the 95% confidence interval around the mean did not go below 4). While it might be suggested that participants were “straightlining” their surveys, analysis of responses did not indicate such behavior. Instead, due to the participants’ academic concentration and enrollment in a highly-competitive university, it seems likely that the sample was biased toward students who tend to have higher motivation and fairly effective learning strategies. In short, while students varied significantly, they did so only in a partial range of the surveys and their integrated academic performance was much higher than the average learner (i.e., even if they used different mixtures of strategies, those strategies “worked”). If we had survey results from USMA, we might see similar results due to high motivation and effective learning strategies. A useful future study would be to include a broader population to identify differences in use and benefit from system features.

That said, two student characteristics still predicted performance and learning. Self-reported lack of anxiety predicted higher knowledge test scores, slightly for the pre-test (r=0.24; p<0.05) and moderately for the post-test (r=0.38; p<0.001). The mechanism for this effect is not entirely clear, but may represent greater comfort and expertise in traditional test-taking, particularly since knowledge check questions were shallow true-false items. Self-reported MSLQ organization (e.g., organizing key class concepts) was likewise correlated with higher learning on situational judgment test items. Particularly for scenario-based training, skills to organize knowledge might be an important factor. Due to the richer sources of information such as the dialogue, virtual character reactions, and shifting conversational skills needed, mentally identifying the important elements and organizing them may lead to more effective learning. This aligns with prior research on other scenario-based ITS, such as Crystal Island, which uses a virtual notebook to help students organize their knowledge (Rowe et al., 2011) and also with research in ITS such as AutoTutor-3D, which found that only a subset of learners appeared to be able to systematically explore and benefit from the simulation (Graesser, Chipman, Haynes, & Olney, 2005). This also indicates that even among higher-performing students, certain strategies are not necessarily universal and need support (e.g., organization and a calm approach to test-taking).

Adaptive support for the scenario (the coach) when added to the intrinsic scenario responses also had a few notable results. First, hints made learners more likely to select a correct response on the next choice in both the non-military and USMA populations. For the non-military data set with pre-test/post-test scores, we noticed that hints made scenario performance a slightly weaker predictor of post-test scores. However, there was little evidence that these hints led to greater learning gains (learning from both conditions was nearly equivalent, despite significant differences in hint prevalence). As such, these hints might often be redundant with the in-scenario feedback by the character. This implies that the scenario-based ITS may not need as much explicit real-time guidance, if the scenario itself provides sufficient feedback and context. In terms of the delayed guidance of the AAR, this is an area for future work since the non-military population's use of the AAR was limited, and the USMA data does not have pre-test and post-test scores.

One finding for the non-military population was their overconfidence in their skills, despite low self-rated experience. Confidence appeared to increase when more time was spent viewing the AAR after repeating Being Heard, which allowed them to see their improvement. However, post-test scores were not correlated with confidence, so this additional time spent with the AAR was not necessarily productive (e.g., identifying mistakes to remedy) so much as possibly reviewing their improvement. This high level of overconfidence may explain the limited use of the AAR.

In the USMA behavioral data, a number of gender differences were seen suggesting that females were better performers (e.g., on average: quicker to respond, higher scores). One could hypothesize a cultural difference between the males and females in this population (e.g., females had more practice with relevant skills such as active listening outside of school). However, more investigation is needed to explain why significant differences were not seen in the non-military population (e.g., the USMA data is more than four times larger so more non-military data is needed, or the males and females of the non-military population are more culturally homogenous).

Long term, a goal of this research is to understand how different students benefit from ITS adaptation, both in terms of engagement and learning. However, one question raised by this research has been: What kind of engagement do we want? For example, when both the scenario and a coach provide feedback, which should the learner engage with? This raises the larger issue of first-order engagement (e.g., attention, time on task) vs. second-order engagement (e.g., regulating attention, including disengaging from unproductive learning activities). Future research might be able to look at this issue by intentionally varying the value of learning activities to identify when and how learners are able to properly regulate their attention and learning. This would help find the skills and cues that effectively-engaged learners use to benefit from a system, which could then be encouraged.

ACKNOWLEDGEMENTS

The effort described here is supported by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. Thanks to the ELITE/INOTS and Engage teams for the use of their software and content. At USMA, we thank the Military Leadership, Engineering Psychology, and West Point Simulation Center leadership teams for agreeing to support our study and readily providing access to all the necessary resources.

REFERENCES

Baker, R. S., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, I., Wagner, A. Z., . . . Beck, J. E. (2006). Adapting to when students game an intelligent tutoring system. In *Proc. of the 8th International Conference on Intelligent Tutoring Systems (ITS).*

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*, 223-241.

Beck, J. E. (2005). Engagement tracing: using response times to model student disengagement. In *Proc. of the 12th International Conference on Artificial Intelligence in Education (AIED).*

Chi, M., Jordan, P., & VanLehn, K. (2014). When is tutorial dialogue more effective than step-based tutoring? In *Proc. of the 12th International Conference on Intelligent Tutoring Systems (ITS).*

Christenson, S. L., Reschly, A. L., & Wylie, C. (Eds.). (2012). *Handbook of research on student engagement*, New York: Springer.

D'Mello, S., & Graesser, A. (2012a). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS) - Special issue on highlights of the decade in interactive intelligent systems, 2* (4), 1-38.

D'Mello, S., & Graesser, A. (2012b). Dynamics of affective states during complex learning. *Learning and Instruction, 22* (2), 145-157.

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies, 70*, 377-398.

Dweck, C. S. (2006). *Mindset: The new psychology of success*, Random House.

Graesser, A. C. (2009). Inaugural editorial for journal of educational psychology. *Journal of Educational Psychology, 101* (2), 259-261.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48* (4), 612-618.

Hays, M. J., Campbell, J. C., Trimmer, M. A., Poore, J. C., Webb, A. K., & King, T. K. (2012). Can role-play with virtual humans teach interpersonal skills? In *Proc. of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).*

Jackson, G. T., Graesser, A. C., & McNamara, D. (2009). What students expect may have more impact than what they know or feel. In *Proc. of the 14th International Conference on Artificial Intelligence in Education (AIED).*

Kim, J. M., Randall W. Hill, J., Durlach, P. J., Lane, H. C., Forbell, E., Core, M., . . . Hart, J. (2009). BiLAT: A game-based environment for practicing negotiation in a cultural context. *International Journal of Artificial Intelligence in Education, Special Issue on Ill-Defined Domains, 19* (3), 289-308.

Lehman, B., D'Mello, S. K., Strain, A. C., Gross, M., Dobbins, A., Wallace, P., . . . Graesser, A. C. (2011). Inducing and tracking confusion with contradictions during critical thinking and scientific reasoning. In *Proc. of the 15th International Conference on Artificial Intelligence in Education (AIED).*

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37*, 91-106.

Pintrich, P. R., Smith, D. A., García, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and psychological measurement, 53* (3), 801-813.

Rodrigo, M. M. T., Baker, R. S. J. D., & Rossi, L. (2013). Student off-task behavior in computer-based learning in the Philippines: Comparison to prior research in the USA. *Teachers College Record, 115* (10), 1-27.

Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education, 21* (2), 115-133.

van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proc. of the 16th Innovative Applications of Artificial Intelligence Conference (IAAI).*

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a united view. *MIS Quarterly, 27* (3), 425-478.

Wixon, M., Baker, R. S., Gobert, J., Ocumpaugh, J., & Bachmann, M. (2012). WTF? detecting students who are conducting inquiry without thinking fastidiously. In *Proc. of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP).*