AD_____

Award Number: W81XWH-16-1-0460

TITLE: Applied Cognitive Models of Behavior and Errors Patterns

PRINCIPAL INVESTIGATOR: Dr. Robert Wray

CONTRACTING ORGANIZATION: Soar Technology, Inc.
Ann Arbor, Mi 48105

REPORT DATE: December 2018

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
Dec 2018	Final	15 Aug 2016 - 14 Aug 2018

4. TITLE AND SUBTITLE
Applied Cognitive Models of Behavior and Errors Patterns (SimMarkers)

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-16-1-0460

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Robert E. Wray, PhD.
Kim Stowers, PhD.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Soar Technology, Inc. (SoarTech)
3600 Green Court, Suite 600
Ann Arbor, MI 48105-2588

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report summarizes concept development, technology prototyping and a study focused on improving training tools for medical readiness. The effort had two primary aims. First, develop models that express indicative patterns. An indicative pattern is a pattern of learner behavior that is predictive/indicative of learner state (e.g., "correct" and error behaviors). Second, evaluate the ability of the models to recognize indicative patterns in training scenarios, especially at pivotal opportunities. A pivotal opportunity is a point at which a learner decision/choice results in different outcomes/directions in the training. We describe the pursuit of three complementary tasks in pursuit of these aims: 1) developing medical training scenarios that support detection of indicative patterns for learner decisions, 2) evaluating indicators that may be apt candidates for supporting medical training, and 3) implementing various methods/processes that could be used to capture indicative patterns in a manner that is both inexpensive and non-intrusive to the learner. This report summarizes progress and accomplishment toward both aims over the entire period of performance.

15. SUBJECT TERMS
computer-based learning, adaptive learning, behavioral patterns, emergency medical technician.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
Unclassified	Unclassified	Unclassified	None	39	19b. TELEPHONE NUMBER *(Include Area Code)*

TABLE OF CONTENTS

1. INTRODUCTION

This report summarizes concept development, technology prototyping and verification and validation studies focused on improving training tools for medical readiness. The effort had two primary aims. The first aim was to develop models that express indicative patterns. An indicative pattern is a recurring feature (pattern) of learner behavior that can be used to predict or indicate learner state. For example, in this effort, we focused on attempting to identify patterns of behavior (primarily, mouse movements) that could provide insight on the level of confidence of a "correct" or erroneous learner decision. The second aim was to evaluate the ability of these models to recognize indicative patterns in training scenarios, especially at pivotal opportunities. A pivotal opportunity is a point at which a learner decision/choice results in different outcomes/directions in terms of the training. In this effort, we focused on the use of adaptive remediation for learner choices. When a learner makes a choice, the feedback that is provided to the learner is modulated by imputed patterns of learner state, based on observations of the system as the learner was making the decision. We hypothesize that the impact of more targeted remediation based on such patterns can improve learning.

The report describes the pursuit of three complementary tasks in pursuit of these aims:
1) developing medical training scenarios that support detection of indicative patterns for learner decisions, 2) evaluating specific indicators that may be apt candidates for supporting medical training, and 3) implementing various methods/processes that could be used to capture indicative patterns in a manner that is both inexpensive and non-intrusive to the learner. This report summarizes progress and accomplishment toward both aims over the entire period of performance. Results from the human-subjects study show modest but statistically significant learning gains from subjects who experienced adaptive, marker-modulated remediation in comparison to the learning gains of subjects who experienced non-adaptive remediation. Further, analyses of mouse movements captured during the experiment suggest that future marker-based adaptation can be tuned to specific, recognizable patterns of mouse movements. This improved precision in markers and subsequent tailoring to the learner during interaction with the learning system suggest the potential for even greater impacts on learning using more precise markers.

2. KEYWORDS

Computer-based learning, adaptive learning, behavioral patterns, emergency medical technician (EMT), mouse-tracking, behavioral indicators

3. OVERALL PROJECT SUMMARY:

Personalized learning, in which a learning environment adapts to the abilities, needs, and preferences of individual learners, has been identified as a "Grand Challenge" for 21st century research and engineering [1]. The benefits of adaptive learning environments include more efficient learning [2], improved attention and motivation [3], the development of less rigid and more flexible decision making [4], and improved transfer of learning to settings in which learned knowledge is used and applied [5-7].

Improved and personalized learning has particular application for more pervasive and less costly medical training, which often is delivered primarily by human instructors in classes with modest student-to-teacher ratios. Human instruction and mentoring is very valuable and desirable, but adaptive personalization methods offer an opportunity to deliver effective introductory and basic training, thus potentially enabling a single human instructor to train many more students by better preparing them for coaching and instruction from experts.

Adaptation to a learner usually requires a model of the learner that is frequently updated as a learner progresses through a curriculum. Creating a complete and accurate learner model is difficult, however. Markers are designed to improve learner modeling. The model of the learner is frequently updated as a learner progresses through a curriculum [8]. The targeting of adaptive techniques, such as scaffolding [9] and competency matching [10, 11] depends on the accuracy (and, to some degree, precision) of the learner model. When the model better reflects the learner's actual knowledge, skills, and attitudes at any point during the learning, the targeting of the adaptive method to the learner generally improves [10]. Creating a complete and accurate learner model is difficult, however. In addition to estimating learner capability from formal and informal assessment within the environment [12-15], researchers have explored many behavioral, physiological, and even neurological indicators or "markers" that can provide additional context for estimating a learner's cognitive state and improving the dynamic assessment of the learner.

Behavioral sensors (posture, eye trackers), physiological sensors (Galvanic skin response), and neurological sensors (EEG) have all been used to assess and track learner arousal/attention in learning environments [16]. These sensors provide details information but at the cost of introducing uncommon and costly new hardware requirements for the learning environment. However, there is significant and growing scientific evidence that the temporal patterns of mouse movements during selection tasks can provide reliable insight into the cognitive state of subjects [17, 18]. Mouse-based markers may be noisier (less diagnostically precise) than neuro-cognitive markers associated with specialized sensors but they are omnipresent on standard computer workstations where many learning environments are deployed. Thus, this effort focused on evaluating the impact of the behavioral markers on the adaptive learning system to improve learning outcomes, taking into account the noise and uncertainty of measure inherent in unspecialized sources.

This focus commonplace hardware to make behavioral observations, such as a computer mouse, distinguishes this effort from work that uses more specialized sensors to recognize indicative patterns. The study we conducted was designed to provide insights into the potential benefits

(and limitations) of using behavioral patterns derived from everyday and pervasive hardware to improve learning outcomes for medical training. The results, while modest and somewhat equivocal, suggest that there is value in capturing and encoding models of these patterns. This work thus offers a foundation for on-going and new learning applications that use models of behavioral patterns to improve learner assessment and targeted of learning content based on those improved assessments.

The statement of work for the effort is summarized in Table 1, including a short description of each major task. Note that Tasks 1 and 2 are focused on specific aim 1 (present pivotal opportunities and elicit indicative patterns) and Task 3 is focused on specific aim 2 (develop computational models of the patterns). In the following, we discuss Objectives, Results, Progress and Accomplishments for each task listed in the Statement of Work.

Table 1. Project Statement of Work

Task 1. Scenario Development
This task is to develop and to validate training content and scenarios. Scenarios are implemented within the Adaptive Perceptual and Cognitive Training System (APACTS). Training scenarios are designed to include supportive, constructive guidance and feedback to present when the learner takes any given action—both for acceptable responses and for erroneous ones. Scenarios are focused on scene size-up for Emergency Medical Technicians. These scenarios involve healthcare content appropriate for an entry-level learner to become familiar with, with a variety of situations portrayed across the entire set of scenarios.
Task 2. Study Design and Data Collection
This task primary focus is to design a study to test the effectiveness of scenarios in identifying behavioral and error patterns in the learning environment and to then conduct the study, collecting and analyzing the resulting data. As an initial step in study design, this task includes an analytic study designed to estimate parameters important for the eventual study design such as the required accuracy of behavioral markers to support effective adaptation for learning based on indicative patterns.
Task 3. Process Modeling
This task is to create models of participant behaviors across the scenarios developed in Task 1. The models compare acceptable behaviors (such as the correct answer to a direct question) and the indicative patterns that led to a chosen answer (such as the mouse movements and dwell times associated with the choice). The models are also developed to be integrated estimates of proficiency and checks on learning (such as explicit questions). At each pivotal opportunity, where a participant is to make a decision in the scenario, we will extend APACTS to record the participant's actions along with the time, and form an assessment against one or more learning objectives. The resulting history of estimates over performance in the scenario can provide insights into the specific progress of learning.

Task 1: Scenario Development

The task was focused on the development of training scenarios that included variation (alternative paths) to support the use of markers and evaluation of those markers in effort studies. Specific objectives and results were to:

Objectives and Results

1. Identify sources of training materials.
   - **This objective was met**. After an initial search for candidate medical training content was unsuccessful, we developed medical training scenarios for the emergency medical technician (EMT) domain in house. One of the advantages of the EMT domain is that it offered a standardized curriculum on which to build training scenarios.
2. Develop instructional design for the scenarios.
   - **This objective was met.** We developed both a complete instructional design and a basic instructional template for each training scenario, drawing directly from the standardized EMT curriculum.
3. Assess and validate the instructional design.
   - **This objective was met**. The standardized, national curriculum has been previously validated. The scenarios developed for this effort hew closely in content with the standard curriculum. We also consulted with an EMT subject matter expert to review specific content presentations, focusing especially on images and content revisions based on the expert recommendations.
4. Implement the instructional design in APACTS.
   - **This objective was met.** We implemented five distinct training units/scenarios in APACTS: 1) A generalized introduction to "scene size up" (a specific EMT task), 2) identification of general hazards, 3-5) recognizing and assessing specific features of hazards, mechanisms of injury, and injuries resulting from head-on collisions, side-impact collisions, and rear-end collisions. In addition, the pre/post-test used in the study was also implemented in APACTS, enabling subjects in the study to complete all experimental steps within a unified software environment.
5. Encode domain meta-data (learning objectives, expected error types, etc.) in APACTS scenarios
   - **This objective was met.** We extended the APACTS learning environment to support the requirements for responding to behavioral patterns and encoded the learning objectives from the standard curriculum into the APACTS scenarios. Learning objectives were encoded to correspond to each individual curriculum unit, as outlined above.

Progress and Accomplishments with Discussion

After search and evaluation of potential options for content, we chose the Emergency Medical Technician (EMT) training and, in particular, one unit within that training. EMT training has several advantages for the effort: 1) curriculum requirements are standardized [19], which essentially places some bounds on the role of instructional design within content design; 2) many organizations offer EMT courses and there are many resources on the web about EMT training, which has alleviated some of content-generation constraints and the need for specialized expertise (i.e., in comparison to combat medics) for creation and validation; and 3) EMT programs (including the subset we have chosen) require development of cognitive, perceptual,

and psychomotor skill. In the effort, we focused on the first two of these, but having more than one type of skill that needs to be developed should help demonstrate the value of behavioral markers for differentiating learning needs.

We chose to focus on the "Scene Size-up" component within the EMT course. The recommended time for this lesson in the standardized curriculum is 1 hour. Within this lesson there are cognitive, affective, and psychomotor learning goals and the goals include not only gaining knowledge but also being able to demonstrate and apply that knowledge during the course of the lesson. The relatively short duration of the lesson with a relatively wide variety of learning objectives and types of objectives, made it a reasonable choice for testing the development of markers, because adaptive choices can potentially focus on choosing alternatives among these categories rather than fine-grained distinctions within a few learning objectives.

The instructional design for the study included the following units:

- Introduction (What is scene size up?)
    - Key Concepts (Introduce terms such as mechanism of injury (MOI)
    - Assessing the complexity of the scene
- Identifying Hazards (general introduction)
- Vehicle Injuries (general intro)
    - Front-end collision (conditions, unique hazards, MOIs, injuries)
    - Side-impact collision (conditions, unique hazards, MOIs, injuries)
    - Rear-end collision (conditions, unique hazards, MOIs, injuries)

A requirement for the scientific goals of this effort was to enable the learner to receive different kinds of information/task different paths based on the decisions the learner made at these *pivotal opportunities*. However, we also faced a tension that content development is expensive and time-consuming and enabling many different and varying paths in the learning content would impact both project goals (resources devoted to content development) and experimental control for any formal evaluation. For example, any unique learning path could result in variation of time on task and result in too few learners exploring that path (given resource constraints on the size of the study) to provide statistical significance.

To enable both some variation in path while also enabling sufficient study control, we developed an approach to scenario development that allows variation that immediately "folds back" to control the amount of variation, a technique our team has deployed in previous simulation-training systems [20, 21]. For each curriculum unit/lesson, we developed an overall template, the structure of which is summarized in Figure 1. Each unit includes some number of introductory "frames" (comparable to a briefing slide) that introduces the topic, terms, and provides examples and explanations. The learner is then presented with a series of vignettes that require a decision/choices. These are the pivotal opportunities in the instructional presentation. The content/frame that the learner is presented next depends on the choice the learner makes. There are generally five distinct choices at each pivotal opportunity defined in the template:

- Move on to the next item (which could another pivotal opportunity or new content)
- Reconsider answer / repeat

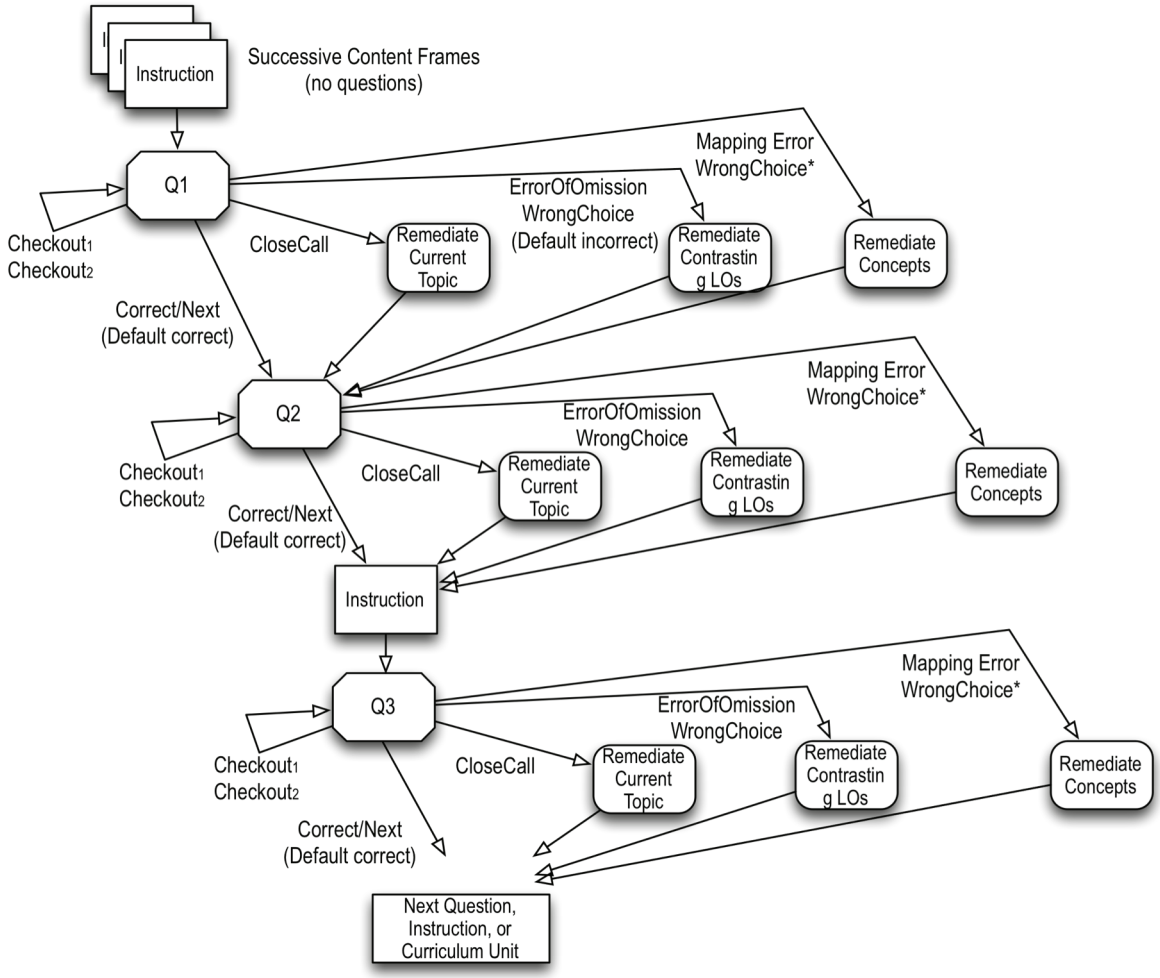


Figure 1. The basic structure of APACTS EMT Scenarios in support of the study.

- Remediate current topic: Feedback is provided that is focused on the current topic and relatively fine-grained distinctions about the topic.
- Remediate contrasting learning objectives: Feedback is provided that discusses differences between the current topic and/or learning objective and another one present in the question choices (e.g., evaluating potential mechanisms of injury between side-impact and rear-end collisions)
- Remediate concepts: Feedback is provided that focuses on high-level conceptual distinctions, such as the difference between a mechanism of injury (the physical forces that can result in patterns of injury) and the injury itself.

Figure 2 illustrates an example question/pivotal opportunity from the hazards lesson and lists the remediation text associated with the responses according to the categories above. *The remediate current topic* feedback focuses on distinguishing the conditions under which a specific perimeter value should be established. The *remediate contrasting learning objective* feedback contrasts the safety perimeter step with the related step of deciding whether backup is needed. The *remediate concept* feedback is comparatively less specific and puts the safety perimeter assessment into the overall context of scene size up.

RCT: The blue car has spilled some liquid, likely fuel given the location at the rear of the car. When there is no fire or fuel spill, choose a perimeter of 50 ft. When there is, set a larger, 100 ft. perimeter for zone.

RLO: The best next step in this scene is to establish a safety perimeter of 100 ft. (100 ft. because of the fuel spill.) Additional backup is likely not needed for this scene.

RCpt: The best next step in this scene is to establish a safety perimeter of 100 ft. (100 ft. because of the fuel spill.) Beginning any evaluation and treatment of patients should start only after you have completed the initial scene up and taken steps to secure the scene.

Figure 2. Example Pivotal Opportunity with Various Remediation Offerings.

For this effort, these choices were not hard-wired to specific learner responses. Instead, the system uses the computational models of behavioral patterns (discussed further under Task 3) to evaluate which content option is most apt for the current situation. Table 2 (next page) summarizes how the patterns of behavior, learner response, and content requirements interact in the system. (As we discuss further below, not all of these choices were included in the study, but the capabilities were developed within the software.) The kinds of learner responses and learner errors called out, as well as recommended interventions are drawn from various empirical validations in the learning sciences, as cited in the table. The decision context for pedagogical remediation using indicative patterns thus takes into account not only the learner's specific decision/response, but also current estimates of skill for the learning objectives relevant to the decision, and the behavioral markers observed (The Task 3 presentation describes the specific algorithm and the choices it makes).

Examples of implemented scenarios are included in Appendix B, the IRB protocol for the primary study. We used photo manipulation tools to insert actors into photographic scenes presented to subjects (see Figure 3 in the next section for an example).

Table 2. Varying responses to learner decisions.

Name	Description	Content Requirements	Pedagogical Response
Checkout1	Learner makes a rapid, high confidence choice before there is sufficient time to interpret the content of the questions.	Estimate of the time it takes to comprehend situation in the frame and/or read an individual question.	Repeat the presentation of the question (no advance until credible attempt).
Checkout2	Learner chooses a "checkout" option from the question options.	Questions/decisions posed to the user include choices that are obviously incorrect/distractor.	Repeat the presentation of the question (no advance until credible attempt). Coach feedback could be included longer term ("c'mon now…").
Mapping Error [22]	Student makes a selection that indicates a misunderstanding of the problem (an example of a categorization error).	Remediation that supports clarifying concept/category differences [23].	Present mapping error remediation whenever the user makes a mapping error and there is mapping-error remediation available.
Error of Omission [22, 24]	For questions that require recognizing multiple correct Reponses, learner chooses a response that leaves out one of the correct choices.	Requires remediation focused on recognizing all relevant features to the decision.	Present remediation and identify specific choice relative to all available choices.
Wrong Choice (Error of Commission) [22] [24]	The user chooses an incorrect response from a multiple-choice question.	Requires remediation of specific options within the question.	Depends on pattern identified. If mouse tracking is able to distinguish between another (correct) choice the user considered and this one, offer remediation for distinguishing between the choices. If user appears to treat all options equally (confused, little evident confident in choice, re-present the decision).
CloseCall1	User weighs several options including the correct one, but ends up making a wrong choice.	Remediation of the question focused on fine-grained distinctions (detailed and specific feedback) [23].	Remediate the question. If mouse tracking is able to distinguish between evaluation of chosen (incorrect) choice and correct one, offer remediation focused on the choices.
LuckyGuess	The user chooses a correct answer but demonstrates (via patterns) little confidence in that answer.	Remediation of the question.	Remediate the question (even though choice was correct). Explain why the choice was correct. Treat as closecall1 if mouse tracking is able to distinguish between another (incorrect) choice the user considered and this one, offer remediation for distinguishing between the choices.

Task 2: Study Design and Data Collection

The task was focused on the evaluation of the process models that produced markers, with a focus on the impact of those markers on learning. Specific objectives and results were to:

Objectives and Results

1. Design and conduct an analytic (verification) study to inform the design of a human-subjects (validation) study.
   - **This objective was met.** The verification study is summarized in [25], which is attached as Appendix A. This analysis enabled us to estimate learning impacts across a large space of learning design alternatives. The results of this analysis lead to us to understand that the study required a larger number of content options for each pivotal opportunity than originally planned and that the study would require a larger number of subjects (75-100) than the original, notional plan (about 50 subjects). As discussed in the previous section, we attempted to balance additional content development requirements, as well as the need for a larger study, within the overall aims of the effort.
2. Design a human-subjects study with the goal of investigating the impact of behavioral markers in an adaptive learning environment.
   - **This objective was met.** The approved protocol for the human-subjects validation study is included in this report as Appendix B.
3. Prepare formal documentation for the study, submit to Institutional Review Board, and obtain approvals from IRB and Army HRPO to conduct the study.
   - **This objective was met.** The study protocol documented in Appendix B received IRB approval on 21 Jul 2017. HRPO required a small change in the protocol and the IRB-approved, amended protocol was approved by HRPO on 22 November 2017. Continuing approval (after the first year) was obtained on 9 Jul 2018.
4. Conduct the study (including subject recruitment, data collection, etc.).
   - **This objective was partially met.** We planned to conduct the primary data collection during the Spring of 2018 but delays in approvals for a contract modification resulted in the subject recruitment not beginning in earnest until the Spring term has concluded. We collected usable data on 62 subjects, rather than the planned 100 subjects. No significant problems were encountered in running the study or data collection (i.e., instrumentation worked as designed) and the subject data collected was sufficient to obtain statistical significance for a primary study hypothesis.
5. Perform data analysis on collected data and summarize overall results and recommendations.
   - **This objective was partially met.** Overall data analysis has been completed and is summarized in this document. Further data analysis can be conducted as well, especially synthesizing more fine-grained models of mouse movements informed by individual subject mouse data and using that analysis to inform refinement of the system interpretation of those models for more precise adaptation.

Progress and Accomplishments with Discussion

(a) *Verification Study*

The goal of the verification-study design was to establish reasonable bounds on potential learning benefits for indicators in an adaptive training context. The study builds on prior work

establishing the use of verification methodologies for the preliminary evaluation of adaptive training systems [26, 27].

The study employed a simulated students paradigm [28-32] to assess theoretical benefits of more targeted assessment via indicative patterns. A secondary goal of the verification study was to identify an appropriate region(s) along a learning curve for human studies. For example, it may be useful to focus more on intermediate or advanced learners to see a large difference in outcomes than novice learners. These kinds of issues reflect why waiting to design the human subjects study until after the verification study is completed is preferable. The primary results of the study were:

- *Behavioral markers must be highly accurate* to facilitate observable impacts on learning given basic constraints on the study design. This outcome led us to focus optimizing mouse-tracking before investigating other sources of behavioral markers, as mouse-tracking has been shown to be fairly reliable in many realistic usage contexts [33].
- A relatively *large number of alternatives are needed at each pivotal opportunity* to effect observable changes in learning outcomes. The content design takes this factor into account in two distinct ways:
    (1) We planned to increase the number of content alternatives available at each pivotal opportunity. This change would have resulted in much larger investment in content; however, the verification study shows that having just a few choices at each opportunity is not sufficient for discrimination across the number of pivotal opportunities a learner could complete in 60-90m of learning experience. As discussed above, we attempted to strike a balance in the implementation of the scenarios between the outcomes of the verification study and the resource limitations of the effort.
    (2) We designed each pivotal opportunity so that the learner faces choices that correspond to a small number of learning objectives (2 or 3) rather than any learning objective in the curriculum. This approach imposes more constraint on content development, but ensures that the resulting feedback is targeted to the learner's misconceptions when incorrect or suboptimal choices are made.
- Behavioral markers will have *greater impact and discrimination for novice learners*. Given study constraints, the impacts of behavioral markers will be more much evident (discriminable from the resulting data) if the learners are not already knowledgeable of the domain. This result led us to focus on a more general target population for the study (college students) than a population already familiar medical procedures like medical or nursing students.

A more complete presentation of the verification study is summarized in a conference paper presented and published during this effort [25], included in this report as Appendix A.

(b) ***Validation/Human-Subjects Study***

Based on the verification study, we designed the human subjects study and documented a protocol for that study. The research compares the results of learning between a medical learning unit presented in a non-adaptive (fixed) sequence to that same content presented in an adaptive sequence that focuses on targeted remediation. As above, the curriculum units focus on "Scene

Size Up," a required curriculum component used in Emergency Medical Technician (EMT) training [19]. These units (both adaptive and non-adaptive) were presented to the subject population in order to assess the utility of markers to improve adaptive learning in emergency medical environments.

The following variables of interest were implemented and observed in the study:

- **Instructional approach**: The overall instructional approach of the learning environment. For this study, there were distinct instructional approaches:
    - Non-adaptive/traditional: An instructional unit that is presented in a fixed sequence to all learners. All learner responses receive the same sequence and remediation.
    - Adaptive based on performance (only): An instructional unit in which specific content presentations are constructed/chosen based on learner performance and subsequent estimates of learner knowledge and skill.
    - Adaptive based on performance and markers: An instructional unit that is dynamically constructed/chosen based on a combination of direct learner observation (as above) and behavior markers.
- **Markers**: Patterns of observed behavior that are hypothesized to have a role in improving a learner model.
- **Knowledge gain**: A measure of the post-test performance of subjects, relative to pre-test performance.

This study was implemented as a between-subjects design, with "instructional approach" being the independent variable of interest. Instructional approach was manipulated at three levels (as discussed above): non-adaptive, adaptive based on performance (only) and adaptive based on performance and markers.

The primary dependent variable was "knowledge gain", as measured by difference scores between pre- and post-tests given to participants. The pre- and post-tests consisted of 26 identical questions of equal weight. Questions were scored as correct or incorrect. Questions in which a subject indicated one choice when several choices were apt received partial credit for the answer based on the total number of correct choices. An example of a pre-/post-test question is illustrated as Figure 3.

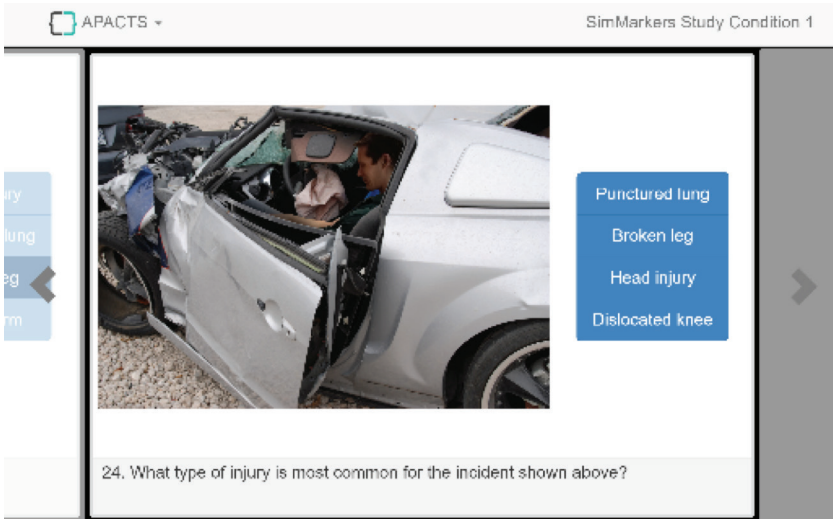


Figure 3. Screen shot of a pre-/post-test question in APACTS.

Additionally, behavioral markers derived from dynamic tracking of mouse movements, were used to predict learner needs and adapt the learning environment by presenting remediation targeted to actual responses. The combination of these variables enabled the study to address the primary hypotheses, as well as quantify the utility of the chosen adaptive learning models for improving learning in medical environments.

A total of 67 subjects were recruited and 61 subjects completed the study successfully.[1] The overall results of the study in terms of the differences in pre- and post-test scores are summarized in Table 3 and normalized knowledge gain [34] between pre- and post-test scores is computed in the right column from the scores. These data are illustrated in Figure 4, with the vertical axis representing the percentage score (out of 26) rather than the raw scores in the table. Error bars represent one standard deviation on either side of the mean.

Table 3. High-level Summary of Primary Study Variables

	Subjects	Pre-test Avg.	Post-test Avg.	Knowledge Gain
Non-adaptive	22	15.33	18.70	3.9%
Adaptive	21	15.12	19.91	5.4%
Adaptive with markers	18	15.69	20.56	5.7%
Total	61	15.37	19.67	

A paired-sample t-test was completed to assess the difference in scores between pre and post-data. Based on the analyses, there is evidence ($t = 13.866$, $p < .001$) that the tutoring intervention as a whole improves learning outcomes. This finding does not take into account differences between tutoring strategies implemented. However, it does suggest that the tutoring program as a whole was effective for helping novice learners grow in their knowledge of EMT practices, specifically scene size-up. Results from an ANOVA (next section) can indicate whether a particular method of tutoring was more effective.

A one-way ANOVA with post-hoc (Tukey) test was completed to assess differences between conditions. This analysis was completed for both pre- and post-test scores, though the intention was to examine

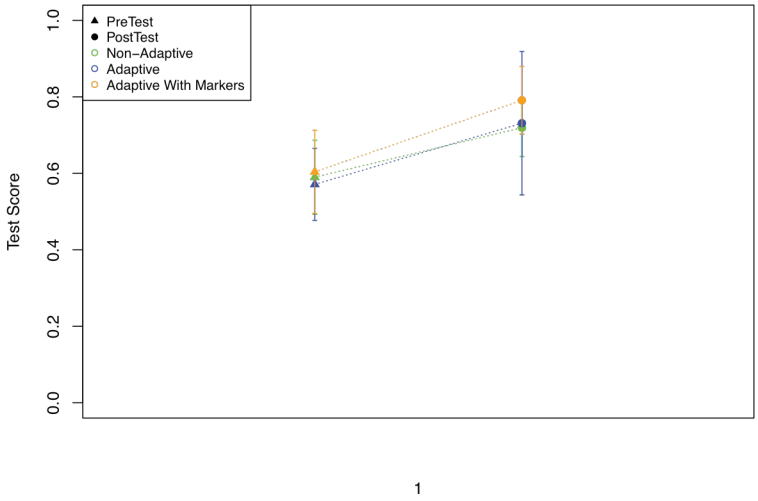


Figure 4. Comparison of pre- and post-test average scores for the three experimental conditions in the study.

[1] We gathered data on 62 subjects but one subject's post-test data was either corrupted or not completed by the subject (post-test score = 0) and is not included in this summary data.

differences between post-test scores between conditions. No statistically significant difference was observed between pre-test conditions (as would be expected). Demographic data collected from subjects generally demonstrated little/no specific EMT knowledge. However, as the pre-test scores indicate, general knowledge of vehicle accidents and hazards allowed subjects to score more highly than we targeted/anticipated.

For post-test scores, a significant difference is detected between groups ($F(2, 60) = 3.63$, $p < .05$, partial $\eta^2 = .11$). A Tukey post-hoc test revealed a significant difference ($p < .05$) between scores in condition 1 (M = 15.34, SD = 2.52, 95% CI [14.22, 16.45]) and condition 3 (M = 15.70, SD = 2.82, 95% CI[14.29, 17.10]). In particular, the mean difference between scores in these conditions was 1.86 (95% CI [.15, 3.57]), with participants scoring on average higher in condition 3.

This result suggests that, for this study, the adaptive tutoring strategy based on both performance and behavioral markers (mouse movements) was more effective than not adapting tutoring at all. However, it is unclear if adapting tutoring based on performance alone was helpful (this finding was not significant). Aside from the indication of significance in difference between non-adaptive and fully adaptive (based on performance + markers), the effect size was also medium in nature, giving some additional support for this finding. In general, the combination of higher than expected pre-test scores and fewer subjects generally resulted in less clear differences in the conditions than hypothesized, although the results we did obtain show a significant difference in outcomes obtained with adaption and behavioral markers in comparison to the traditional/standard computer-based training approach.

The study suggested that the mouse-tracking-based markers piloted and developed aided adaptive selection of content targeted to the learner. The process models that generate the predictive markers were developed via analysis of general mouse tracking behavior, algorithm design and development based on this analysis, and initial piloting that helped define parameters needed for the models, such as an estimate of the time needed to read questions of particular length. The Task 3 description and Appendix C further describe the development of these models. Following data collection in the study, we sought to further understand the patterns of mouse tracking observed in the learning scenarios, as well as to identify opportunities to improve these models based on more substantive data collection than was feasible in the piloting.

Data analysis of mouse tracking focused on two questions:

1. **Comparisons of aggregate and individual mousing behavior.** The process models developed prior to the study were based on general mouse behaviors. For example, we developed pilot questions of significantly different (character/word number) lengths in order to estimate the distribution of the length of time subjects might need to read a question before answering it. The aggregate analysis focuses on coming to a similar understanding for specific questions used in the study as well as suggesting techniques for identifying patterns of response based on user interaction with the questions.

   Figure 5 compares the performance of the population on two pre-test questions (4 and 6) and those same questions on the post-test. The mouse tracking behavior is divided into three categories of behavior: void (yellow; time before the mouse is placed over a

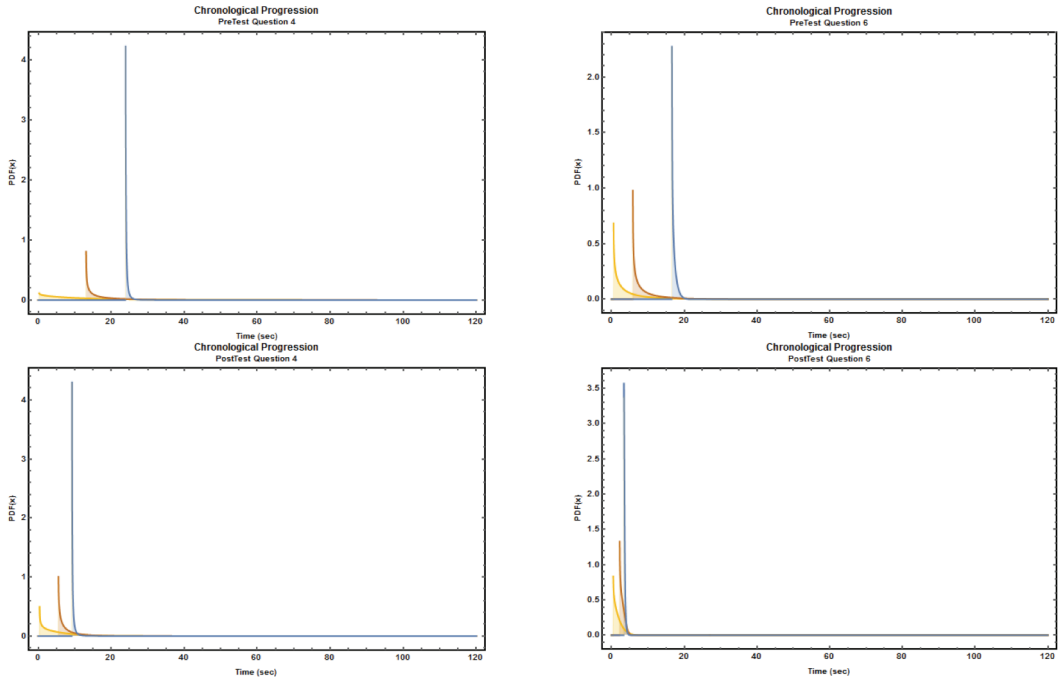


Figure 5. Comparison of aggregate subject mousing behavior for questions 4(left) and 6 (right) for pre-test (top) and post-test (below).

question response), hover (brown; time spent hovering over responses), and click-to-response (purplish-blue; the time from the final click to submission of the answer). The charts plot a gamma distribution of all subject responses for the question for those three variables as well as the total time (light blue line) with the y-axis representing the percentage of users that would respond in the timeframe. (The gamma distribution was the best-fit distribution over all questions and subjects and the four categories of time tracked for the study.) The hover and click/submit distributions are linearly shifted based on the standard deviation of the previous step. These plots thus provide a gross summary of the behavior of all subjects on a question.

Roughly, we expected that for many users, void time represents reading the question and some consideration of responses; hover time represents deliberate consideration of responses; and click-to-response is potentially a surrogate for confidence in an answer. (For now, we ignore multiple clicks on answers; see next for more discussion of that issue.) In the figure, we can observe that for both questions, void time is reduced between pre- and post-test (perhaps indicating some familiarity with the question), but total hover time and click-to-response are shifted further to the left (less time), suggesting that, on the whole, subjects required less time to evaluate choices and their choices were finalized more quickly and perhaps decisively.

In particular, by comparing aggregate and individual performance on pre-test questions, in comparison to post-test questions, we can begin to estimate the relative difficulty of questions based on these patterns of responses. In the figure, pre-test question 4 was answered correctly by 30/62 subjects (48%) and performance was almost unchanged in

the post-test (35 correct; 56%). Comparing pre- and post-test mousing behavior in the charts, there is less void time but the distributions of the responses in hover and click/submit time are almost the same, suggesting the relatively similar performance on this question for the pre- and post-test. In contrast, Question 6, the pre-test performance is comparable (35 correct, 56%), but post-test performance increases significantly (59 correct, 95%). The highly correct response rate is evident in the plot in comparison to both the pre-test performance on that same question and in comparison to the post-test question that was not as easy to answer.

2. **Identification of learner and question sub-patterns**. We anticipated that subject data would enable us to identify additional features and patterns in mousing behavior. For example, we expected some users would have relatively large void times and little hover time (perhaps not moving the mouse at all during reading and thinking) while others would be more demonstrative mousers. If we can identify these patterns post-hoc, then we can, in future work, refine the process models that generate the markers based on more individualized patterns of observed mousing behavior.

The mouse tracking data did show this specific pattern, among others. Figure 6 shows an example from Pre-test Question 11 where this different pattern is visible directly in the data. The chart plots the total time each user in the study spent on the question with the first movement to the mouse over a question as t=0. The light brown bars to the left of 0 seconds is then hover time, the dark brown is the aggregate time spent mousing over item responses, and the purple bars represent the time between the final click for a selected item until submission. Questions are listed from top to bottom in the descending order of the hover period.

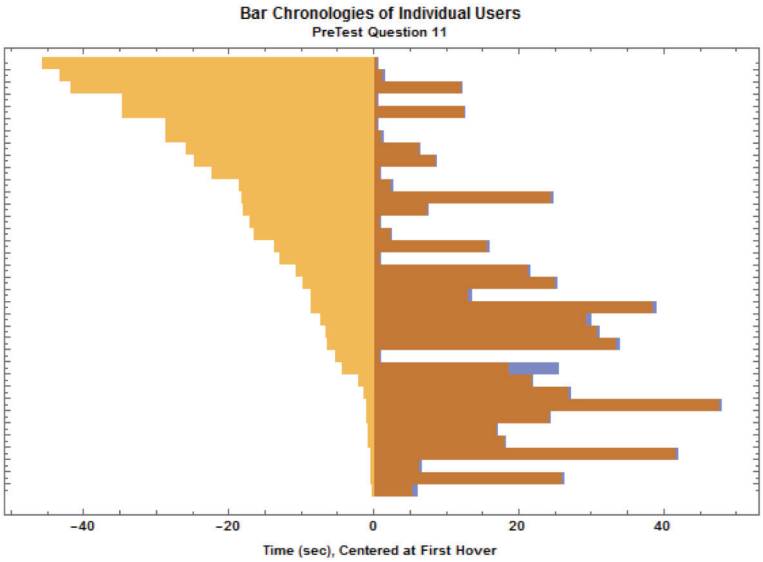


Figure 6. Illustrative example of three distinct patterns of mouse movements.

In this example, there are three clusters of responses. A little less than half the users spent most of the time reading (> 15s) with little mouse movement until a response was generated. About a fourth of the users spent 5-10 seconds reading the question and then hovering over responses, and then about a third spent very little time (< 5s) on the question before starting to mouse over responses. Not all responses in the data were so clearly visible, but clustering of the patterns of responses to questions bore out this trend generally over the pre-test and unit questions. (The post-test clustering was less

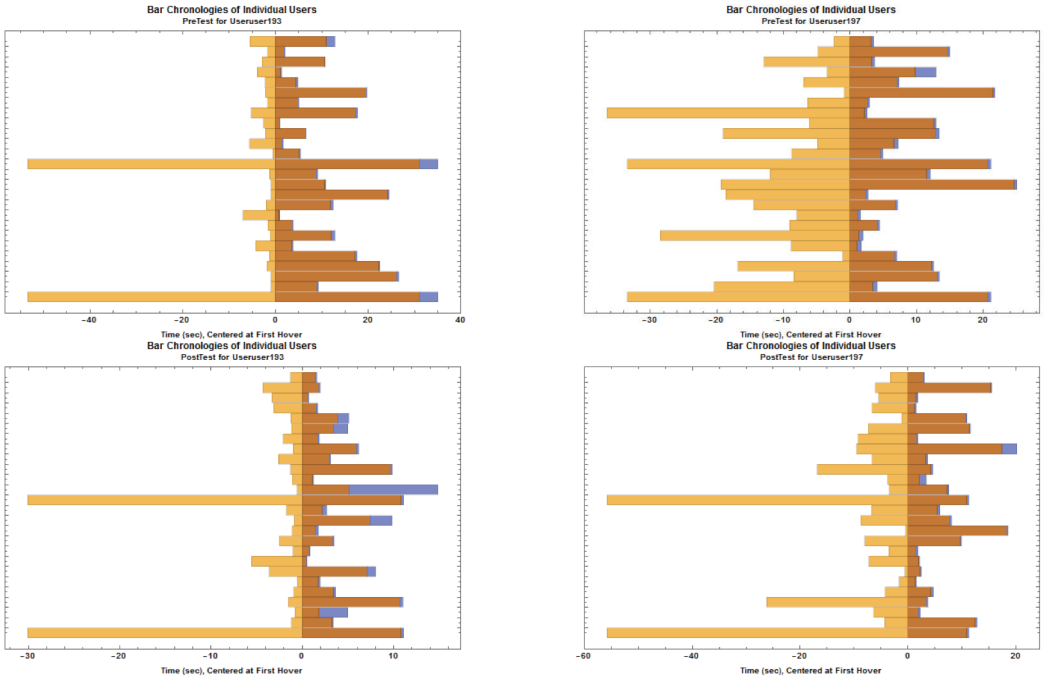


Figure 7. Comparing individual mousing behavior between pre-test and post-test.

consistent in part due to the significantly less time overall spent on post-test questions by the users.)

Somewhat surprisingly, although we expected subjects in the "hover over responses as they read" would generally submit their questions well after subjects in the other two clusters. However, this turned out to not be the case, or was masked by another effect. Instead, the length of time spent between click and submission was more of a function of the individual subjects (some subjects consistently paused before submission in what we attribute to a "check your work" deliberateness) and the difficulty of the question. This pattern is suggested in Figure 7, which compares the void/hover/click-and-submit behavior of two users for pre- and post-test questions.

Further analysis of the pattern of movement in hovering over answers also can provide insights that could be used to improve marker precision. Consider the four examples illustrated in Figure 8. Each of these charts shows the movement of the mouse over the time spent for each subject on Question 11 (note the absolute length of the x-axis changes for the different plots). Void time (mouse not on the question or a response) is represented at y=0 and any spent hovering over the questions (Q) is represented at the top of the plot. Each item response (in this case, 1-7) is represented as an integer on the y-axis.

The left column illustrates two users who did not move the mouse over a question item for a significant amount of time while the right column shows tow subjects who began mousing over responses almost immediately. These columns correspond to two of the groups discussed above. However, there are also apparent patterns in the rows of these examples as well. The subjects in the top row are mousing quickly over multiple item

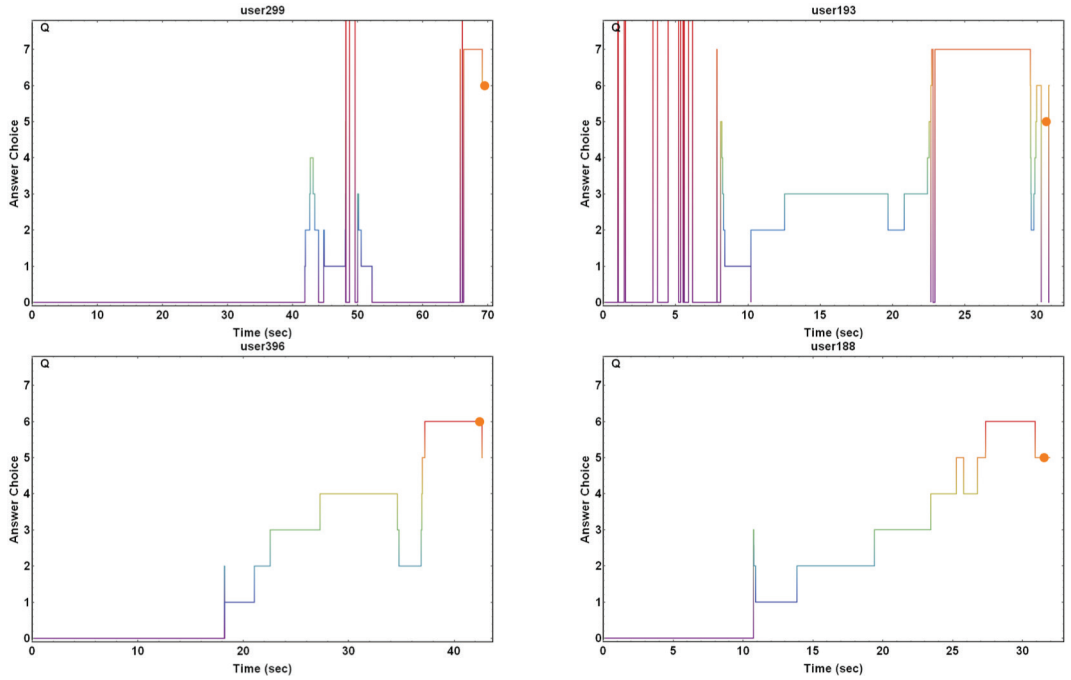


Figure 8. "Indecisive" (top) and "Systematic" (bottom) patterns of mouse movement between the "read first" (left) and "hover first" groups.

responses (more quickly than the responses can be read). This pattern of movement is suggestive that the locus of their attention is on the item responses, but not any particular one. We term this pattern "indecisive" because there is no identifiable consideration of individual choices. In the second row, the movement of the mousing is (generally) more regular and systematic with a clear progression through the choices evident in the mousing. This pattern was much more common in the "hover immediately" subjects, but also present in the "read first" group as well.

One of the advantages of this kind of pattern recognition in mousing would be to refine the confidence estimation that was used from the mouse tracking to improve the precision of the markers. For example, the "lucky guess" marker uses a ratio of time spent over "reasonable responses" in comparison to "non-reasonable" responses to attempt to estimate when a user may have guessed correctly. In the case of the user who mouses over all responses before making a choice, the way this marker is determined would be improved by taking into account the recognition of the particular user's pattern of mousing over items.

The complete study protocol is included as Appendix B. A conference paper summarizing motivations and study design [35] is included in the report as Appendix D. A more thorough presentation of the analysis of the outcomes from analysis of pre- and post-test results is presented in Appendix E.

Task 3: Process Modeling

Objectives and Results

1. Assess modeling options and develop a framework of indicators.
   - **This objective was met.** We evaluated options and identified mouse tracking as the behavioral indicator of highest priority given study constraints.
2. Define an algorithmic approach for assigning meaning to behavior indicators in the context of the learning environment and interactions among learning objectives.
   - **This objective was met.** Building from general frameworks for characterizing learning and misconceptions (e.g., *Mind Bugs*) and previous work reifying learning concepts in a practical software implementation, we created a method for assigning meaning/interpretation to patterns of mouse movements and mouse behaviors.
3. Develop models for mouse tracking (primary modeling option).
   - **This objective was met.** Drawing from the results from the previous two objectives, we have implemented, tested, and verified computational models that perform the interpretation of mouse tracking, recognizing learner patterns and assigning them an interpretation in the context of the current learning situation.
4. Integrate the models in the APACTS learning environment.
   - **This objective was met.** The models developed under the previous objective have been integrated within the APACTS software for use in APACTS learning environments. This integration included software testing and verification of software functionality of the models within units of learning content. In addition, after the initial proof-of-concept integration and verification was accomplished, the mouse-tracking capabilities were fully integrated into APACTS and can now be used by other researchers and application developers using APACTS.
5. Refine and extend models.
   - **This objective was not met.** We conducted some piloting and model refinement based on piloting. However, because data collection extended to the end of the project for the main study, there is not sufficient time on the effort to refine the models. Several recommendations for further refinement based on data analysis are evident (and outlined below).

Progress and Accomplishments with Discussion

We reviewed and evaluated two existing approaches to behavior and error classification: Van Lehn's learner-behavior classification scheme [22] and Rasmussen's Skills, Rules and Knowledge [36]. After evaluation of each of these methods and reference to them in the design of the verification study, we determined to use Van Lehn's *Mind Bugs* taxonomy for classification of errors. This taxonomy is more comprehensive than SRK and while it is also more descriptive than SRK (i.e., rather than generative), we did not identify any major stumbling blocks in encoding recognition rules from the taxonomy in the error recognition system. We have recently extended the framework to include the Knowledge-Learning-Instruction (KLI) [37] and the Interactive, Constructive, Active, and Passive (ICAP) [38] frameworks. These frameworks take a more up-to-date and comprehensive view of learners and learning environments and have facilitated making more fine-grained distinctions in assessment and task contexts for modeling learner behaviors and errors.

For encoding recognizers or "markers" in the learning environment, we developed models that built on a prior constraint-based behavior modeling system [39] to encode non-symbolic behavior patterns. We focused primarily on mouse movements and mousing behavior generally as an indicator of both cognitive and affective state. Patterns of mouse movements have reasonable correlation with a learner's affective state [40] and multiple studies suggest that learner mouse movements can be effective in identifying learner cognitive state [17, 18, 41].

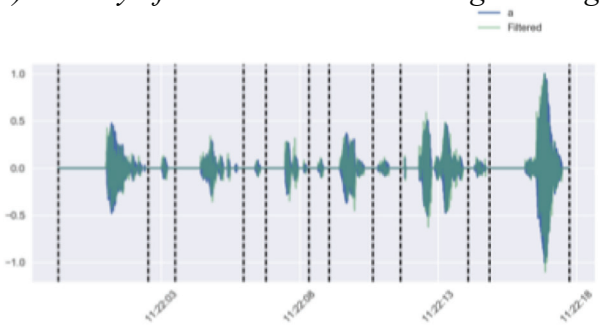
Figure 9 summarizes the mouse-tracking algorithm, which perform the first step in the recognition process. The learner has been asked to "annotate" the image in the APACTS frame, identifying any objects in the image that is a "hazard" as defined in the EMT curriculum. Positional information is captured, along with the velocity and acceleration of the mouse movement and mouse clicks (represented in the diagrams by the vertical, dashed lines). The velocity and acceleration graphs include examples of both raw (blue) and filtered (green) data. Filters help reduce some of the noise due to inadvertent mouse movements and mouse jitter.

The positions of key objects in the scene are labeled as meta-data (labeled in the scenario development accomplished in Task 1; illustrated in Figure 10), enabling the mouse-tracking algorithm to relate mouse actions to learner activity. For example, in the first and second mouse click events (2nd and 3rd vertical lines in the figures), these areas are associated with the bystanders/potential patients in front of the cars. Although the behaviors appear quite different (compare the two velocity spikes), these are readily classified as comparable outcomes in the learning environment via

(a) tracking learner mouse movements

(b) (x,y) position of movement

(c) velocity of mouse movement during tracking



(d) acceleration of mouse movement

Figure 9. Basic steps in tracking mouse movement.

the use of the labeled areas in the content illustrations.

By application of Fitt's Law [42] and the filtered data, the tracking algorithms are used to estimate the confidence of an individual decision. For example, in the latter part of the scenario, the mouse tracks to a few locations but the user does not make a mouse click. By comparison of velocities and accelerations of these different movement patterns, the algorithm attempts to assess the confidence of the learner's decision.

A more complete presentation of the mouse-tracking algorithms is summarized in a conference paper presented and published during this effort [43], included in this report as Appendix C.

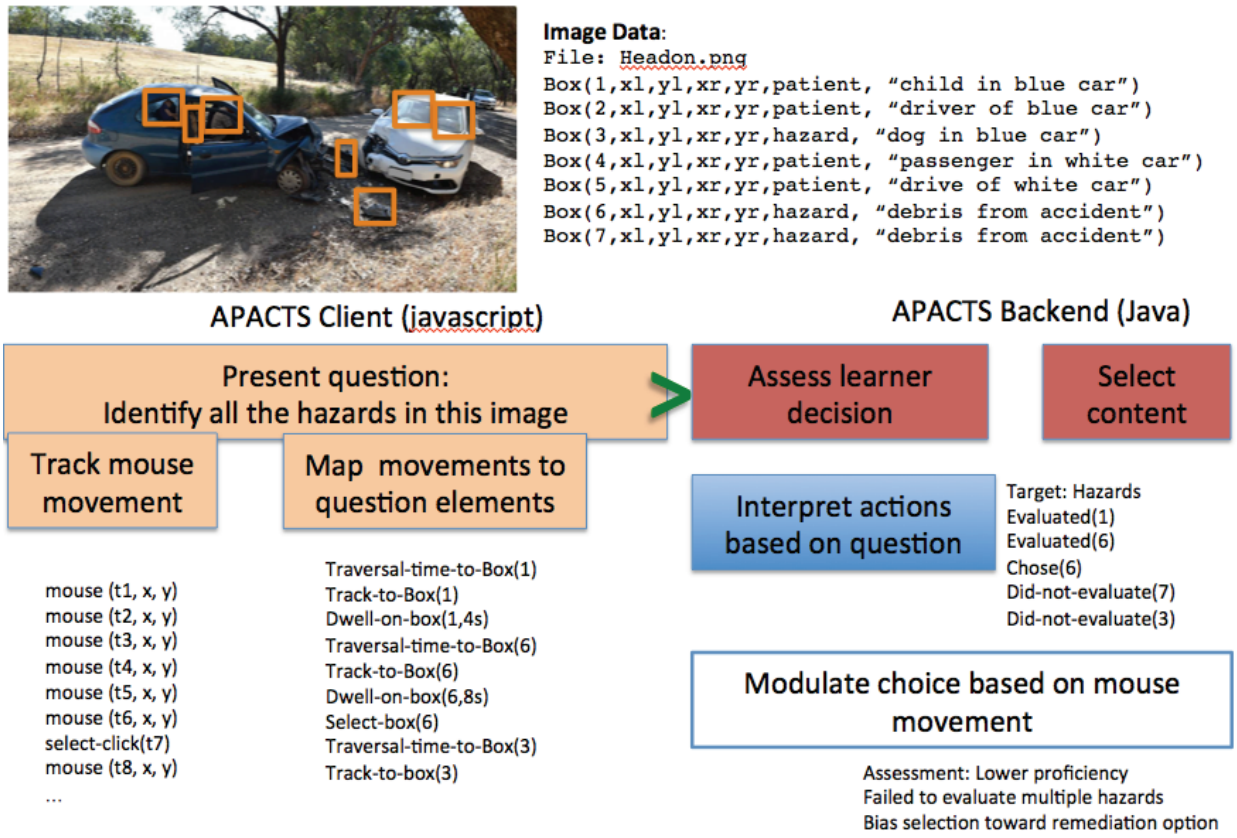


Figure 10. Translating Mouse Movement into Learner Assessment.

Figure 10 summarizes the information flow that results in the mouse-tracking assessments. Following the low-level tracking illustrated in Figure 9 (summarized in the "track mouse movement" component in the figure above), the captured movements are mapped to task interpretations, such as moving to a labeled object ("track to box"), dwelling on a box, and a normalized traversal time. The mouse tracking feeds the primary model (blue component), which focuses on the interpretation and evaluation of the learner's choices. In this example, the model is indicating which of the labeled areas were evaluated by the learner, which of those boxes the learner actually chose, and which boxes the learner did not appear to evaluate based on mouse movements.

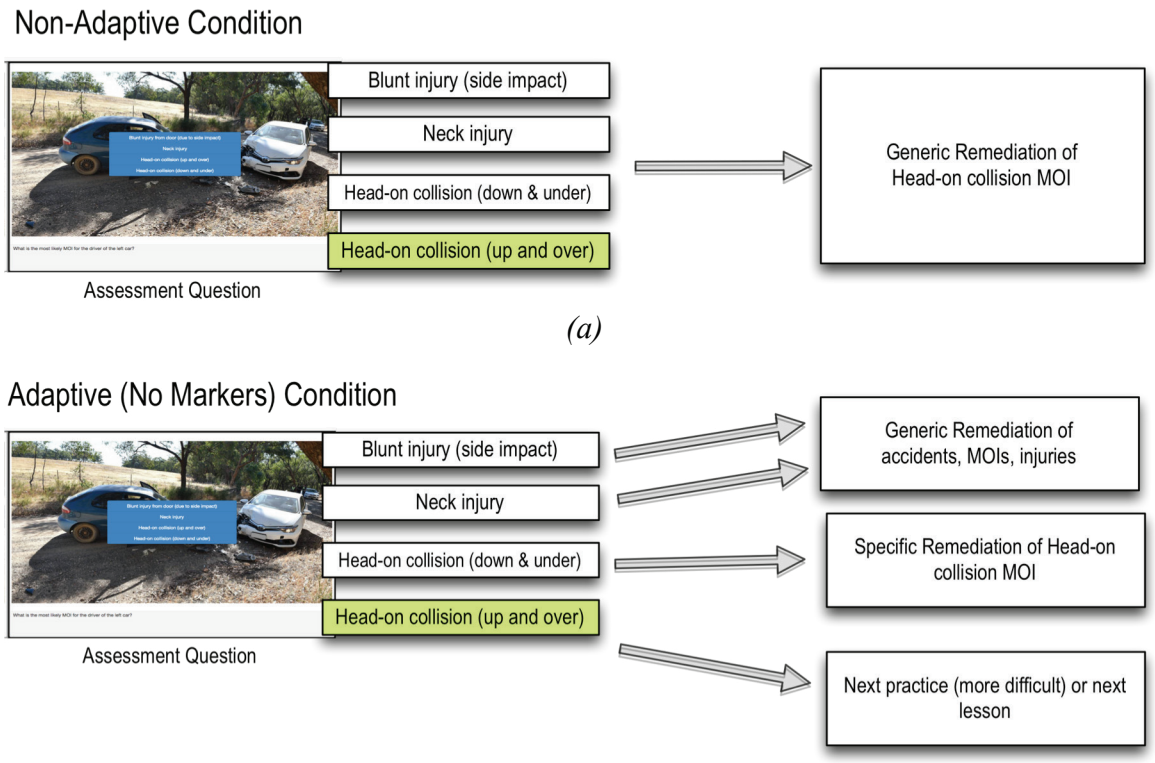


Figure 11. Non-adaptive (a) and adaptive content selection (b and c). In (c), markers derived from mouse-tracking enable a choice of alternative tailoring /remediation for the same learner answer.

These evaluations then feed to the content selection algorithm in APACTS, which determines what content the learner sees next. In the situation illustrated in Figure 10, the learner's proficiency estimate for relevant learning objectives is low and the mouse tracking lets the system understand that the learner did not appear to evaluate hazards in the image. The lack of evaluation results in a bias toward one of the remediation options.

Figure 11 contrasts the way the model impacts the final content selection decision in the APACTS system in comparison to non-adaptive content presentation and adaptation based solely

on a learner's choice. In this multiple-choice question example, the learner is asked to classify the mechanism of injury (MOI). In the non-adaptive case (a), the learner just receives general feedback about their response. In the simple adaptive case, the feedback that the learner receives is conditioned on the learner's response. For example, if the learner chooses an answer that is "close" to the right answer, such as confusing the two patterns of MOIs common in head-on collisions, the feedback that is provided is highly specific to the head-on collision MOI. If the learner instead chooses a response that is not consistent with a head-on collision (like the first choice), a more general remediation is offered that is intended to remediate differences between learning objectives (e.g., side-impact collisions vs. head-on collisions).

The assessment of mouse movements can enable different tailoring responses for the same learner choice. For example, if the learner spends a lot time evaluating all of the options (including item (b), which is a different category of response than the others), the system will choose to remediate MOIs vs. injuries even though the learner's eventual response was the correct choice. In this case, the models provide additional context for interpreting learner activity and tailoring the presentation of content to the learner.

At the implementation level, the learner choice influences both the skill estimates of the learner the presentation of specific remediation options. As discussed above, the costs of content development limited the individual content choices we could construct under this effort. As a consequence, the decision making component of APACTS, the Pedagogical Director [44] and its learner model [45] were adapted to support the additional information that the mouse-tracking algorithms and process models could provide.

For remediation, the range of choice that the Pedagogical Director can make is limited to those choices called out in Figure 1. To summarize:

- Next item (choose the subsequent item in the curriculum. Standard choice.)
- Remediate current topic (more information/emphasis of learning points)
    - Example: More information about head-on collisions.
- Remediate differences between current topic and related topic
    - Example: Distinguishing between head-on and side-impact collisions or Distinguishing between head-on and rear-end collisions.
- Remediate conceptual and terminological issues
    - Example: Explain differences between mechanism of injury and injury
- Repeat (present the same item again)

For the study content, we ensured that every option was available for all choices, but we designed the software so that not every one of these options needs to available for each content-selection decision that the Pedagogical Director makes.

The learner model is built on TrueSkill©, a model designed by Microsoft originally for use in adaptive computer gaming. Soar Technology has adopted TrueSkill for adaptive learning systems and integrated it into APACTS to allow APACTS to develop estimates of the likelihood of particular learners, at particular points in time in their learning trajectory, answering specific questions correctly [45]. For the mouse-tracking models and study, we fixed the estimated difficulty of questions, with the result that the learner model is producing an estimate of

likelihood of correct answer for a given question as well as the variance of the estimate. (For the study, the same question could be presented multiple times to the same learner but we did not change the computed estimate based on re-presentation of the question.). The learner model estimation can be summarized by the following equation:

$$E_{l,o} = T(observations_{l,o})$$

That is, the estimate E of the learner l's likelihood of answering a given question about learning objective o, is a function T of the previous observations of the learner answering questions for that learning objective. The function T is simply the TrueSkill algorithm with a fixed question difficulty and reduces to $T_n = T(observation_n, T_{n-1})$.

The estimate and its variance range over (0…100%) and can vary continuously as the system gets more observations of the learner. However, for the purposes of a controlled study, we needed to limit the variation introduced by the continuous variable E, which would make comparison of experiences between various subjects more difficult. Instead, we discretized E into the following six categories:

- Unlikely (low chance of correct answer; 0-33%)/low variance
- Unlikely/high variance
- Equivocal (34-67%)/low variance
- Equivocal (34-67%)/high variance
- Likely (> 67%)/low variance
- Likely/high variance

We use the likelihood of the estimate as a proxy for "skill level" and the inverse of the variance as a proxy for "confidence" in the decision making process. For example, $E(s_{high},c_{low})$ means that True Skill estimates the learner's ability to answer correctly to be > 67% for this question (likely to answer correctly) but we have "low" confidence in that assessment because the variance is large. Using this categorization, the Pedagogical Director chooses specific content to display following an incorrect (or very fast, correct response), based on the decision matrix summarized in Table 4.

As summarized in the table, this approach allows the system to condition specific content presentation choices based on the learner's responses and the estimates of skill and confidence in those estimates. As an example, a very fast response that is produced in about the time it takes to simply read the question (Checkout1), results in a repeat of the question for most learners. The exception is for learners whose skill for this learning objective is estimated to be to high and the system's confidence in that estimate is also high. In that case, the system simply moves on to the next question. This decision reflects the choice that the learner might be able to process and answer faster than a less skilled learner (and also that the cost of repeating the question to the high skilled learner could be de-motivating).

Even after the reduction of the estimates into distinct categories, the scope of the study limited (number of participants) limited options for making more nuanced mapping decisions because the number of observations for any specific mapping choice was likely to be very limited. For example, we considered various formulations of remediations for certain classes of error, which

would have allowed improved tailoring. A low skill learner might just be presented with basic remediation about the learning objective, a higher skill learner could be presented with more discriminating information related to the error. Varying the level of feedback for a specific question is surmountable from a content authoring perspective, via content generation algorithms [46]. However, even if the content authoring challenges could be resolved, for a study of only 100 subjects, the number of observations for any cell in the table would be no more than 6 (100 subjects / (3 study conditions * 5 LOs)) and in many cases might be only 1 or 0 for many LOs. This led us to limit the variation in the decisions across answer categories (rows in the table) in order to attempt to gather more consistent observations about each kind of error.

Table 4. Mapping skill and confidence to remediation choices

	E(low skill, low conf)	E(s_{low}, c_{high})	E(s_{med}, c_{low})	E(s_{med}, c_{high})	E(s_{high}, c_{low})	E(s_{high}, c_{high})
Checkout1	Repeat	Repeat	Repeat	Repeat	Repeat	Next
Checkout2	Repeat	Repeat	Repeat	Repeat	Repeat	Repeat
CloseCall1	Remediate current topic	Remediate current topic	Remediate current topic	Remediate current topic	Remediate current topic	Remediate current topic
Error of Omission	Remediate differences (identified LO)	Remediate differences (identified LO)	Remediate differences (identified LO)	Remediate differences (identified LO)	Remediate differences (identified LO)	Remediate differences (identified LO)
Lucky Guess	Remediate differences (identified LO) or Remediate concepts (no LO)	Remediate differences (identified LO) or Remediate concepts (no LO)	Remediate differences (identified LO) or Remediate concepts (no LO)	Remediate differences (identified LO) or Remediate concepts (no LO)	Next	Next
Mapping Error	Remediate concepts	Remediate concepts	Remediate concepts	Remediate concepts	Remediate concepts	Remediate concepts
Wrong Choice	Remediate differences (identified LO) or Remediate concepts (no LO)	Remediate differences (identified LO) or Remediate concepts (no LO)	Remediate differences (identified LO) or Remediate concepts (no LO)	Remediate differences (identified LO) or Remediate concepts (no LO)	Remediate differences (identified LO) or Remediate concepts (no LO)	Remediate differences (identified LO) or Remediate concepts (no LO)

Although limited in its application in the study, the overall result from this task is that we developed a set of algorithms and tools that allow a learning system to observe mouse movements from users and to recognize various patterns in those mouse movements ("markers") that can be used to inform pedagogical decision-making. Although the implemented and integrated system focuses solely on mouse movements, we also explored the use of other sensors and markers derived from them, such as eye tracking and facial expressions. One of the advantages of the conceptual integration at the level of various markers indicative of learner state is that sensor fusion and downstream use of the markers could be modulated in the future by relative confidence in the markers. We chose mouse tracking because of its high reliability but, with multiple sensors providing inputs on the same set of markers, long-term the system could earn more confidence in its use of these markers, even when the reliability of any individual sensor might be low.

Regulatory Protocol and Activity Status

(c) ***Human Use Regulatory Protocols***

TOTAL PROTOCOLS: *1 human subject research protocol was required to complete the Statement of Work.*

PROTOCOL(S):

TOTAL PROTOCOLS: 1

PROTOCOL (1 of 1 total):

Protocol: **A-19646**

Title: *Assessing the Role of Behavioral Markers in Adaptive Learning*

Target required for clinical significance: N/A (72 for statistical significance in the study design)

Target approved for clinical significance: N/A

SUBMITTED TO AND APPROVED BY:

- **Submitted to** Ethical & Independent Review Services (Soar Technology IRB).
- **Approved by** Ethical & Independent Review Services, 19 July 2017.
- **Continuance by** Ethical & Independent Review Services, 9 July 2018

STATUS:

We received approval from Ethical & Independent Review Services (Soar Technology's IRB) and from HRPO (22 Nov 2017). Subject recruitment began in Apr 2018. The study enrolled 67 subjects.

(i) Number of subjects recruited/original planned target: 70/100
Number of subjects screened/original planned target: 67/100
Number of patients enrolled/original planned target: 67/100
Number of patients completed/original planned target: 62/100

(ii) Report amendments submitted to the IRB and USAMRMC HRPO for review: Amendment 1A (modification to recruitment flyer required by HRPO) was approved by IRB on 11/9/2017. HRPO approval was based on Amendment 1A of the protocol. Continuance was approved by IRB on 9 Jul 2018; HRPO on 18 Aug 2018.

(iii) Adverse event/unanticipated problems involving risks to subjects or others and actions or plans for mitigation:

No unanticipated problems involving risks to subjects occurred during the course of the study.

(b) Use of Human Cadavers for Research Development Test & Evaluation (RDT&E), Education or Training

TOTAL ACTIVITIES: *"No RDT&E, education or training activities involving human cadavers was performed to complete the Statement of Work (SOW)."*

(c) Animal Use Regulatory Protocols

TOTAL PROTOCOL(S): *No animal use research was performed to complete the Statement of Work.*

4. IMPACT

What was the impact on the development of the principal discipline(s) of the project?

The largest potential impact of the specific outcomes of this work is its innovative use of verification analysis to support study design. A practical constraint in the design and development of algorithms and tools for personalized learning is the need to design, implement and integrate adaptive algorithms, oftentimes within complex software environments, without the benefit of a priori large-scale user testing. This constraint is particularly acute in complex training environments, such as those used in distributed simulation and virtual training where premature commitment to an approach may take several years and significant cost increases to correct.

This effort contributed to a developing methodology that employs simulated students and software verification methods to attempt to understand the potential benefits of adaptive algorithms and the requirements they impose on students and instructors prior to full-scale development. The verification study results predicted two of the primary results of the study:

1. Significant diversity in content is needed to observe statistically-significant differentiation in learning outcomes during a short learning session; and
2. Tailoring of remediation will have a greater impact for low-skilled learners during short learning sessions due to ceiling and asymptotic effects on higher-skilled learners.

While these observations could have been anticipated in advance, the verification study resulted in quantitative predictions and recommendations for the subsequent study. Enabling quantitative assessment of design choices prior to implementation offers a substantial benefit and resulting impact for the adaptive learning community.

What was the impact on other disciplines?

Nothing to Report.

What was the impact on technology transfer?

The Adaptive Perceptual and Cognitive Training System (APACTS) tool used on this effort is being used by other projects and groups within Soar Technology for learning sciences research and the development of adaptive training applications. For example, APACTS is currently being used to deliver training for small-unit leader in the US Marine Corps and for Navy cyber defense training. The computational process-models have been integrated with APACTS can be used in these and future applications of this software to training applications. By embedding the mouse-tracking algorithms and process models within APACTS, the results of this effort will transfer to other training applications used by the government.

5. CHANGES/PROBLEMS

Changes in approach and reasons for change
There was one significant change in approach and one change that was minor in comparison to the first change but that had implications for the overall goals of the project.

Shift in study population: We originally planned to focus on medical-community trainees, such as nursing students, for the primary study. The original effort had several consultants who had access to students in medical disciplines and we planned to conduct a study with one or more of these populations. However, this plan proved unworkable for both scientific and practical reasons. First, and most importantly, for a learning session of 60-120m, which is what we proposed and had obtained budget to support for the effort, the verification study showed that very little discrimination between learners would be obtained across study conditions for even moderately knowledgeable subjects. To have an opportunity to see learning effects, we needed a less knowledgeable subject population. Secondarily, engaging consultants to inform a study was useful but was not an apt relationship to conduct a study, given both IRB and HRPO requirements. As a consequence, we formally requested a change to the contract and engaged the University of Alabama as a subcontract partner for the effort. The team at the University of Alabama was then able to recruit from a general population of college students and directly conduct collection of subject data from this more general population.

Balancing Content development and Research Goals: We had expected that scenario development (Task 1) would be a relatively small fraction of the effort, in comparison to the effort expended on Tasks 2 and 3. During the first six months of the effort, we attempted to find existing training content that could be imported into APACTS to support the study and reached out to multiple organizations to attempt to identify content without success. In the end, we built a new content unit ourselves. We leveraged existing instructional design and content recommendations for EMTs but the development of new content (along with at least three content options for every learner decision/pivotal opportunity) consumed much more of the effort's resources than originally expected. The need for much greater content development both limited the scope of the study (one recommendation from the verification study was to extend the study to more units, but that was not cost-feasible) and the number of markers and sensors we could investigate under the effort.

Actual or anticipated problems or delays and actions or plans to resolve them
The only significant delay experienced during the effort was the time needed to obtain permission to modify the contract to add the subcontractor and then to execute that subcontract. We expected that to take 1-2 months and instead it took about 6 months. This delay negatively affected our ability to conduct the majority of the study during a typical university term and contributed to not meeting subject recruitment goals.

Changes that had a significant impact on expenditures
The effort was completed within budget. As above, in comparison to the original plan, significantly more effort was expended on content development than planned or preferred.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

Significant changes in use or care of human subjects
There were no significant changes. The protocol was executed as designed and no significant deviations from the protocol were necessary.

Significant changes in use or care of vertebrate animals
Not applicable.

Significant changes in use of biohazards and/or select agents
Not applicable.

6. PRODUCTS:

Publications:

a. Manuscripts presented/published during the period covered by this report resulting from this project:

Wray, R. E., & Stowers, K. (2017). Interactions between Learner Assessment and Content Requirements: A Verification Approach. *Proceedings of the 8th International Conference on Applied Human Factors and Ergonomics (AHFE 2017) and the Affiliated Conferences*, AHFE 2017, Los Angeles.

- Included as Appendix A

Wearne, A., & Wray, R. E. (2018). Exploration of Behavior Markers to Support Adaptive Learning *Lecture Notes in Computer Science: Proceedings of the 2018 Human Computer Interaction International (HCII) Conference*. Las Vegas.

- Included as Appendix C

Stowers, K., Brady, L., Huh, Y., & Wray, R. E. (2018). *Assessing the Role of Behavioral Markers in Adaptive Learning for Emergency Medical Services.* Paper presented at the Proceedings of the 9th International Conference on Applied Human Factors and Ergonomics (AHFE 2018) and the Affiliated Conferences, AHFE 2018, Orlando.

- Included as Appendix D

b. Publications in preparation:

Wray, R. E., Tanaka, A., Stowers, K., Brady, L., & Huh, Y. (2019). Using Mouse-tracking in Adaptive Learning for Emergency Medical Services Lecture Notes in Computer Science: Proceedings of the 2019 Human Computer Interaction International (HCII) Conference. Orlando.

- This paper will describe the hypotheses, study method, and results as summarized in Appendix E.

Technologies/Techniques
The Adaptive Perceptual and Cognitive Training System (APACTS) tool used on this effort is being used by other projects and groups within Soar Technology for learning sciences research and the development of adaptive training applications. The computational process-models (described in Task 3) have been integrated with APACTS will be used in future applications of this software to training applications.

Inventions, patent applications, and/or licenses
Nothing to report.

Other products
The training materials developed for study undertaken in this study could be used to support EMT training in future computer-based training for EMTs. Additionally, the content-development pipeline established on this project could be employed to speed future content development in similar domains. For example, the process by which captured images were enhanced with post-hoc accident victims and bystanders could be used in future training content to produce content more rapidly than animating images or scripting scenes.

7. PARTICIPANTS and COLLABORATING ORGANIZATIONS

The following individuals worked at least a person-month on the effort:

Soar Technology

Name:	Robert Wray
Project Role	Principal Investigator
Researcher Identifier	1906504 (CITI)
Nearest person month worked	4
Contribution to Project	Oversaw content development and testing. Implemented sample lessons. Oversaw piloting and algorithm refinements.

Name:	Adam Wearne
Project Role	Software Engineer
Nearest person month worked	4
Contribution to Project	Designed and implemented process models to support learner tracking and choice-confidence estimates.

Name:	Nick Giranda
Project Role	Software Engineer
Nearest person month worked	2.5
Contribution to Project	Supported development and refinement of APACTS, content adaptation, and mouse tracking to support piloting and finalization of software for the study. Supported deployment test and configuration/reconfiguration.

Name:	Robin McNeil
Project Role	Content Tester/Data Analyst
Nearest person month worked	1.5
Contribution to Project	Developed instructional design and learning content for the study. Supported data analysis.

Name:	Alyssa Tanaka, PhD.
Project Role	Study coordinator/content developer
Researcher Identifier	28104663 (CITI)
Nearest person month worked	1
Contribution to Project	Helped design and develop learning content. Led quality assurance and evaluation of content. Coordinated study execution with University of Alabama and maintained study records.

Name:	Ross Hoehn, PhD.
Project Role	Data scientist
Nearest person month worked	1
Contribution to Project	Performed analysis on mouse tracking data and developed visualization animations that allow analysts to visualize mouse usage during specific questions.

Name:	Alex Crowell
Project Role	Software Engineer
Nearest person month worked	1
Contribution to Project	Assessed implementation and complexity of mouse-tracking algorithms for full integration into APACTS. Fully integrated mouse tracking algorithms into APACTS to enable future use in other applications of APACTS.

University of Alabama

Name:	Kim Stowers, PhD.
Project Role	Principal Investigator, University of Alabama
Researcher Identifier	2863606 (CITI)
Nearest person month worked	2
Contribution to Project	Designed / completed human subjects study questionnaires and implementation; completed analyses and contributed to dissemination of information.

Name:	Lisa Brady
Project Role	Experimenter
Researcher Identifier	5282480 (CITI)
Nearest person month worked	1
Contribution to Project	Ran participants for human subjects study; contributed to analyses and dissemination of information.

Name:	Youjeong Huh
Project Role	Experimenter
Researcher Identifier	24414679 (CITI)
Nearest person month worked	1
Contribution to Project	Ran participants for human subjects study; contributed to analyses and dissemination of information.

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?
Nothing to Report.

What other organizations were involved as partners?

Organization Name: University of Alabama
Location of Organization: Tuscaloosa, Alabama
Partner's contribution to the project (identify one or more)

- The University of Alabama *collaboration* was led by Dr. Kim Stowers. The University of Alabama team was responsible for subject-data collection, including subject recruitment, subject data collection, and post-processing of the data to support overall data analysis. The university provided *facilities*, especially a dedicated space location to conduct the study and supported recruitment (e.g., distribution of the recruitment poster). Soar Technology and University of Alabama met frequently to coordinate execution of the study and Soar Technology staff trained the Alabama team on the use of the software, facilitating *personnel exchanges* between the staff of the two organizations comprising the team.

8. SPECIAL REPORTING REQUIREMENTS

None.

9. REFERENCES:

[1] National Academy of Engineering, "Grand Challenges for Engineering," National Academy of Sciences/National Academy of Engineering, Washington, DC2008.

[2] B. P. Woolf, *Building intelligent interactive tutors: student-centered strategies for revolutionizing e-learning*: Morgan Kaufman, 2008.

[3] S. D. Craig, A. C. Graesser, J. Sullins, and B. Gholson, "Affect and learning: An exploratory look into the role of affect in learning with AutoTutor," *Journal of Educational Media,* vol. 29, pp. 241-250, 2004.

[4] G. Hatano and K. Inagaki, "Practice makes a difference: Design principles for adaptive expertise," in *Annual Meeting of the American Education Research Association*, New Orleans, 2000.

[5] J. D. Bransford and D. L. Schwartz, "Rethinking Transfer: A Simple Proposal With Multiple Implications," in *Review of Research in Education*. vol. 24, A. Iran-Nejad and P. D. Pearson., Eds., ed American Educational Research Association: Washington, DC, 1999.

[6] C. W. Coultas, R. Grossman, and E. Salas, "Design, Delivery, Evaluation, and Transfer of Training Systems," in *Handbook of Human Factors and Ergonomics*, ed: John Wiley & Sons, Inc., 2012.

[7] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, pp. 1345-1359, 2010.

[8] P. Durlach and R. Spain, "Framework for Instructional Technology," in *Advances in Applied Human Modeling and Simulation*, V. G. Duffy, Ed., ed: CRC Press, 2012.

[9] R. D. Pea, "The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity," *The Journal of the Learning Sciences,* vol. 13, pp. 423-451, 2004.

[10] T. Murray and I. Arroyo, "Toward Measuring and Maintaining the Zone of Proximal Development in Adaptive Instructional Systems," presented at the Proceedings of the 6th International Conference on Intelligent Tutoring Systems, 2002.

[11] L. S. Vygotsky, *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press, 1978.

[12] J. A. Anderson, A. T. Corbett, K. Koedinger, and R. Pelletier, "Cognitive Tutors: Lessons Learned," *The Journal of the Learning Sciences,* vol. 4, pp. 167-207, 1995.

[13] P. Dillenbourg and J. Self, "A framework for learner modeling," *Interactive Learning Environments,* vol. 2, pp. 111-137, 1992.

[14] Z. A. Pardos, N. T. Heffernan, B. Anderson, and C. L. Heffernan, "Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks," in *Handbook of Educational Data Mining*, R. Christobal, Ed., ed: CRC Press, 2010, pp. 417-426.

[15] J. T. Folsom-Kovarik, R. E. Wray, and L. Hamel, "Adaptive Assessment in an Instructor-Mediated System," presented at the Artificial Intelligence in Education (AIED), Memphis, 2013.

[16] J. V. Cohn, D. Nicholson, and D. Schmorrow, Eds., *The PSI Handbook of Virtual Environments for Training and Education*. Westport, CT: Praeger Security International, 2008, p.^pp. Pages.

[17] E. Hehman, R. M. Stolier, and J. B. Freeman, "Advanced mouse-tracking analytic techniques for enhancing psychological science," *Group Processes and Intergroup Relations,* vol. 18, pp. 384-401, 2015.

[18] B. Quétard, J. C. Quinton, M. Mermillod, L. Barca, G. Pezzulo, M. Colomb*, et al.*, "Differential effects of visual uncertainty and contextual guidance on perceptual decisions: Evidence from eye and mouse tracking in visual search," *Journal of Vision,* vol. 16, 2016.

[19] United States Department of Transportation and National Highway Traffic Safety Administration, "EMT-Basic: National Standard Curriculum," ed, 1996.

[20] R. E. Wray, J. T. Folsom-Kovarik, A. Woods, and R. M. Jones, "Motivating narrative representation for training cross-cultural interaction," in *Proceedings of the 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015*, ed Las Vegas: Springer, 2015.

[21] J. T. Folsom-Kovarik, A. Woods, and R. E. Wray, "Designing an Authorable Scenario Representation for Instructor Control over Computationally Tailored Narrative in Training," in *Proceedings of the 29th International FLAIRS Conference*, ed Key Largo: AAAI Press, 2016.

[22] K. Van Lehn, *Mind Bugs*. Cambridge, MA: MIT Press, 1990.

[23] V. J. Shute, "Focus on Formative Feedback," *Review of Educational Research,* vol. 78, pp. 153-189, 2008.

[24] R. E. Wray, A. Woods, and H. Priest, "Applying Gaming Principles to Support Evidence-based Instructional Design," presented at the 2012 Interservice/Industry Training, Simulation, and Education Conference, Orlando, 2012.
[25] R. E. Wray and K. Stowers, "Interactions between Learner Assessment and Content Requirements: A Verification Approach," in *Proceedings of the 8th International Conference on Applied Human Factors and Ergonomics (AHFE 2017) and the Affiliated Conferences, AHFE 2017*, Los Angeles, 2017.
[26] R. E. Wray, B. Bachelor, R. M. Jones, and C. Newton, "Bracketing human performance to support automation for workload reduction: A case study," in *Lecture Notes in Computer Science: Proceedings of the Human Computer Interaction International (HCII) Conference*, ed Los Angeles: Springer-Verlag, 2015.
[27] R. E. Wray, A. Woods, J. Haley, and J. T. Folsom-Kovarik, "Evaluating Instructor Configurability for Adaptive Training " in *Proceedings of the 7th International Conference on Applied Human Factors and Ergonomics (AHFE 2016) and the Affiliated Conferences*, ed Orlando: Springer, 2016.
[28] B. MacLaren and K. R. Koedinger, "When and why does mastery learning work: Instructional experiments with ACT-R "SimStudents"," in *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, Berlin, 2002.
[29] C. J. MacLellan, E. Harpstead, R. Patel, and K. R. Koedinger, "The Apprentice Learner Architecture: Closing the loop between learning theory and educational data," in *Proceedings of the 9th International Conference on Educational Data Mining-EDM'16*, 2016.
[30] K. Van Lehn, "Two pseudo-students: Applications of machine learning to formative evaluation," in *Advanced Research on Computers in Education*, R. Lewis and S. Otsuki, Eds., ed Amsterdam: Elsevier, 1991.
[31] N. Matsuda, E. Yarzebinski, V. Keiser, R. Raizada, W. C. William, G. J. Stylianides*, et al.*, "Cognitive anatomy of tutor learning: Lessons learned with SimStudent," *Journal of Educational Psychology,* vol. 105, pp. 1152-1163, 2013.
[32] C. J. MacLellan, K. R. Koedinger, and N. Matsuda, "Authoring Tutors with SimStudent: An Evaluation of Efficiency and Model Quality," in *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, Honolulu, 2014, pp. 551-560.
[33] M. W. Scerbo, F. G. Freeman, P. J. Mikulka, R. Parasuraman, F. Di Nocero, and L. J. Prinzel III, "The efficacy of psychophysiological measures for implementing adaptive technology," National Aeronautics and Space Administration2001.
[34] R. Hake, "Interactive-Engagement Versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses," *American Journal of Physics,* vol. 66, pp. 64-74, 1998.
[35] K. Stowers, L. Brady, Y. Huh, and R. E. Wray, "Assessing the Role of Behavioral Markers in Adaptive Learning for Emergency Medical Services," in *Proceedings of the 9th International Conference on Applied Human Factors and Ergonomics (AHFE 2018) and the Affiliated Conferences, AHFE 2018*, Orlando, 2018.
[36] J. Rasmussen, "Skills, rules, and knowledge: Signals, signs, and symbols, and other distinctions in human performance models.," *IEEE transactions on systems, man, and cybernetics (SMC),* vol. 13, pp. 257-266, 1983.
[37] K. R. Koedinger, A. T. Corbett, and C. Perfetti, "The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning," *Cognitive Science,* vol. 36, pp. 757-798, 2012.
[38] M. T. H. Chi and R. Wylie, "The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes," *Educational Psychologist,* vol. 49, pp. 219-243, 2014.
[39] A. Woods, B. Stensrud, R. E. Wray, J. Haley, and R. M. Jones, "A Constraint-Based Expert Modeling Approach for Ill-Defined Tutoring Domains," in *Proceedings of the Florida Artificial Intelligence Research Society (FLAIRS) Conference*, ed Hollywood, FL: AAAI Press, 2015.
[40] S. Salmeron-Majadas, O. C. Santos, and J. G. Boticario, "An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context," in *18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, Poland, 2014.
[41] U. Demšar and A. Çöltekin, "Quantifying gaze and mouse interactions on spatial visual interfaces with a new movement analytics methodology," *PLoS ONE,* vol. 12, 2017.
[42] P. M. Fitts and M. I. Posner, *Learning and skilled performance in human performance*. Belmont CA: Brock-Cole, 1967.

[43] A. Wearne and R. E. Wray, "Exploration of Behavior Markers to Support Adaptive Learning," in *Lecture Notes in Computer Science: Proceedings of the 2018 Human Computer Interaction International (HCII) Conference*, ed Las Vegas, 2018.
[44] R. Hubal, M. van Lent, J. Wender, B. Lande, and S. Flanagan, "What does it take to train a good stranger?," in *Proceedings of the International Conference on Cross-Cultural Decision Making*, ed Las Vegas: Springer-Verlag, 2015.
[45] C. Kawatsu, R. Hubal, and R. Marinier, "Predicting Students' Decisions in a Training Simulation: A Novel Application of TrueSkill™," *IEEE Transactions on Computational Intelligence and AI in Games,* vol. PP, 2016.
[46] J. T. Folsom-Kovarik, "Developing a Pattern Recognition Structure to Tailor Mid-Lesson Feedback," in *Proceedings of the 5th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym5)*, R. Sottilare, Ed., ed, 2017.

APPENDIX A: Interactions between Learner Assessment and Content Requirements: A Verification Approach

Interactions between Learner Assessment and Content Requirements: A Verification Approach

Robert E. Wray, Kimberly Stowers

[1] Soar Technology, Inc. 3600 Green Court Suite 600, Ann Arbor, MI USA 48105
wray, kimberly.stowers @soartech.com

Abstract. A practical constraint in the design and development of algorithms and tools for personalized learning is the need to implement adaptive algorithms, oftentimes within complex software environments, without the benefit of a priori large-scale user testing. The lack of such testing makes it difficult to ensure that lessons and guidance from design recommendations and prior studies in other domains has been effectively applied in the training application. This paper summarizes efforts toward a testbed to support verification of adaptive training designs. The testbed operationalizes evidence-based guidance from the research literature and simulated students to enable exploration of design space prior to large-scale implementation. The paper motivates the approach with a specific design question, which is to examine trade-offs between the use of behavioral markers to assess proficiency and the resulting training-content requirements to take advantage of the information that such markers provide.

Keywords: training design, adaptive training

1 Introduction

A practical constraint in the design and development of algorithms and tools for personalized learning is the need to design, to implement and to integrate adaptive algorithms, oftentimes within complex software environments, without the benefit of a priori large-scale user testing. User testing can provide evidence of what adaptive methods are more (and less) beneficial within a particular training setting. The most beneficial, specific methods will usually not be fully known in advance; many potential design options may be apt. Knowledge of the research literature and results can be helpful, but best practices for the design of adaptive training in most training contexts is ever-evolving [1, 2].

This constraint is particularly acute in complex training environments, such as those used in distributed simulation and virtual training. The complexity of software integration and limited access to physical devices can result in commitment to a design that turns out to not offer many direct training benefits. Similarly, a chosen approach may offer a significant improvement in learning effectiveness but the target population cannot realize those benefits because their incoming knowledge and skill is not matched to those benefits provided by the system.

When an algorithm or approach turns out to be poorly chosen, it may take several years to develop and implement an alternative approach. This delay has both immediate and longer-term impacts. The immediate cost is the lack of improvements in training that were anticipated by the training developers. A longer-term, more systemic cost is that these failures in execution can impose greater resistance and new barriers for the adoption of adaptive training generally, resulting in the perception that adaptive training methods are not sufficiently mature to deliver the learning benefits that have been observed in more controlled (and, oftentimes, contained) settings.

As researchers interested in developing and fielding effective adaptive training solutions, we have for several years been developing a methodology that employs simulated students and software verification methods to attempt to understand the potential benefits of adaptive algorithms and the requirements they impose on students and instructors prior to full-scale development [3-5]. We introduce a testbed we are developing to enable exploration of design choices and, to illustrate how the testbed can inform specific design choices, summarize a verification study conducted using the methodology. This study reflects the long-term goal to develop methodology and tools that will help designers understand what (adaptive) features are appropriate/needed for their training needs and to estimate the costs/benefits of different design options.

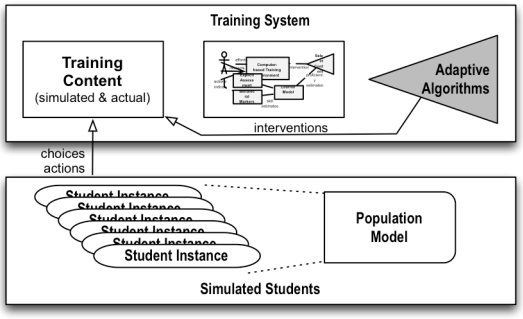


Fig. 1. Conceptual Architecture of Verification Testbed.

2 Testbed for Training Design

Below we briefly introduce the elements of the verification testbed we are developing. The goal of the testbed is to provide a computational tool, with parameters connected to the research literature, that allows a training designer to evaluate assumptions about a design. Fig. 1 illustrates the major components of the testbed and their relationships to one another.

Testbed components are:

1. **Adaptive algorithms**: The testbed typically uses the implementation of adaptive algorithms that would be used in the actual training environment. From a software engineering perspective, this approach allows evaluation and test (or *verification*) of the adaptive solutions within the testbed.
2. **Learning-system architecture**: The learning-system architecture defines how training content will be delivered and the role of adaptive algorithms within the learning environment. We are developing a family of these models for use in the

testbed. The next section introduces the specific model we are using for this analysis (see Fig. 2).

3. **Training content**: The testbed draws on a content repository to deliver training content within the testbed. In some cases, this training content may be the actual content that is to be used in the training application (especially apt when adding adaptive capabilities to an existing training application). In other cases, especially for a new training system being designed, the training content may be simulated.
4. **Simulated students**: The testbed employs simulations or models of students to interact with the training content. The use of simulated students to support training design is becoming more commonplace; some researchers have identified methods to synthesize functional students based on task analyses, cognitive architectures, and machine learning [6, 7]. Analytic tools, such as power law equations, are often also used for modeling learning [8, 9]. The primary requirement for a simulated student is that it provide a response to a learning situation at an appropriate level of abstraction for the simulation of the learning environment.
5. **Population Model**: The population model varies parameters for individual simulated students as they are instantiated. Having a distinct population model (rather than a defined population of simulated students) allows the user of the testbed to explore potential interactions across population assumptions (e.g., students with generally high/low self-efficacy; students generally well-prepared or poorly prepared for the content to be delivered).

Long-term, we envision a flexible and composable software environment that would allow designers to model potential learning designs and evaluate them in a decision analysis aid. Today, we are creating instances of the components illustrated in Fig. 1 to address specific design questions, as discussed next.

3 Motivating Example

As described above, the study we present uses a simulated students paradigm and a simulation of the learning environment to provide quantitative estimates for functional system requirements. The benefit of this approach is that specific learning benefits and the effects of adaptation can be evaluated, at least tentatively, in advance of full-scale implementation. Here we discuss the learning environment being simulated, along with the specific domain we pull learning content from.

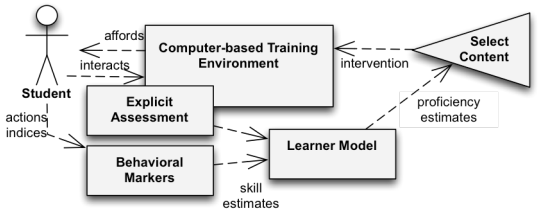


Fig. 2. Model of the learning environment.

Computer-based training (CBT) is actively used across many contexts, including military, medical, and general education. CBTs commonly include didactic instruction

(text and images, audio, and video), opportunities for relatively simple practice, and periodic checks of knowledge. Most CBTs assume a fixed sequence of lessons and may require a student who fails a knowledge check to repeat a lesson. Implementing adaptive training in such a context may yield many benefits, most notably the benefit of accelerating or decelerating the pace at which students move forward in the lesson according to how quickly they are learning, including improved engagement. Adaptive techniques used in CBTs include variable starting points [10], enabling more/less practice [11], hinting and coaching [12, 13], and personalization of content delivery [14, 15].

We are designing and evaluating the role of adaptation in a CBT for Emergency Medical Technician (EMT) certification. EMT courses are offered across the United States, with various states enforcing slightly different requirements. Curriculum is standardized at the US federal level through the National Highway Traffic Safety Administration [16]. This makes EMT training both accessible and applicable. Additionally, EMT certification is a domain of training that can be applied in both national and international civilian and military contexts, making it a highly valuable area for the training improvement. Adaptive training may help streamline the EMT certification process by accommodating learners who may need more or less practice to meet national standards.

For the specific analysis of this paper, we examine a specific lesson in the standard curriculum for EMT training—scene size-up. Scene size-up involves steps taken by an EMT crew when arriving on the scene of an emergency. According to the standard curriculum, in order to develop training within this context, it is necessary to consider the "scene size-up" timeline and specific cognitive, affective, and psychomotor objectives for this task (see table 1). The standard curriculum specifies 9 distinct learning objectives across these three different types of learning objectives.

It would be useful in designing the training environment to have insights and quantitative estimates for the following three questions:

1. *What is the potential size of the learning gain that would be introduced by the use of adaptive methods?* This question sets expectations for the design and helps the designer to understand the relative benefit of adaptive training in the context of the impacts of the full system.
2. *How much unique content is needed to realize the ideal (or at least compelling) learning gains*? Tailoring to the learner typically requires specialized content. If we assume that it is not possible to automate content creation (the typical case), then it would be beneficial to estimate the minimum content needed to realize a (meaningful) gain from adaptive tailoring across the target population.
3. *How accurate do assessment measures need to be to realize (compelling) learning gains?* In order to make adaptive choices, some measurement of the state of the learner during the learning process is typically needed? How accurate do measures need to be to realize the hypothesized gains from adaptive tailoring?

Table 1. Key parameters for the marker/content verification analysis.

Parameter	*Description*	*Study Value(s)*	*Citations*
Base Learning Rate	The learning rate term in a standard power law learning curve (α)	.25	The specific α value is in the range of common values in learning models [8, 9]
Learning Objectives Types	Distinct categories of learning objectives.	3	Cognitive, Affective, Psychomotor from Standard EMT Curriculum [16].
Number of Learning Objectives	Objectives that must be met according to the topic and tasks being learned to complete a scene size-up.	9	9 distinct learning objectives are identified in the standard curriculum [16]
***Z* Score**	A normalized (-1..1) relative match between learner capability and material being presented.	*See text*	This Z-score is an operationalization of the ZPD and is informed by [18] but is adapted to the anticipated training context.
Delta Learning rate	Modification of base learning rate with the assumption that high z-score improves learning rate and low z-score diminishes learning rate.	+/- 25%	This range is comparable to learning gains observed in a similar domain with tailored content matching [15].
Measure Accuracy	The general accuracy of measures used to estimate skill/proficiency.	*See text*	Direct measures can have high accuracy. Indirect measures, such as markers, often can exhibit poor precision and recall.

4. Verification Methodology

To attempt to answer these questions, we developed a simulation of the EMT learning environment within the testbed and developed specific tests to gather data. A summary of the implementation for each testbed component is summarized below. Table 1 lists specific values for some of the primary parameters used in the study. Testbed components:

1. **Adaptive algorithms**: This test focuses on a single adaptive algorithm, which chooses the lesson content that is closest to the estimated proficiency of the learner across all learning objectives. We are interested in the use of other adaptive algorithms, including hinting and coaching. However, in this study, we focus only on lesson selection.
2. **Learning-system architecture**: Modeled as displayed in Fig. 2. We did not distinguish explicit assessment and marker-based measurement, although explicit assessment is generally more accurate than marker-based techniques.
3. **Training content**: We generated several collections of lessons, which are primarily characterized by the target learner profile for the lessons (but not all lessons touch on all learning objectives). The control or baseline lesson condition

is a "progressive" lesson design, which assumes an initial low student proficiency vector and increases target proficiency values (more or less uniformly) across all learning objectives as instruction progresses. This choice is reasonable for most CBTs, although a part-task design is a contrasting option for future study.

4. **Simulated students**: In this design, students were simulated using a power law model. We employed a form of the power law model which computes the impact of a lesson solely from the current lesson and prior learning [17]. This form of the power law allows us to estimate the effect of each individual lesson and assume a heterogeneous collection of lessons. For the study, each "lesson" was estimated to be about 4 minutes of instruction, resulting in 15 distinct lessons (and 14 opportunities for intervention) within the learning design.

   The effect of adaption on learning is estimated by assessing how closely a chosen lesson matches the learner's proficiency profile. The Z(PD)-score is computed as the average mismatch between the lesson (target profile) and student/actual profile for all learning objectives addressed by the lesson. Normalization is applied to the average error to bound to the range [-1...1], where a 1 represents a perfect match and a -1 represents a (near-perfect) mismatch. How precise targeting needs to be is obviously of interest to the adaptive training community. We chose a conservative approach, assuming a functional relationship in which the maximum Z-score rapidly decreases for relatively small targeting errors. In other words, unless targeting is very good, its effect on the learning rate for that lesson will be small.

5. **Population Model**: The primary population variable used in the study is the initial proficiency profile of students. An initial proficiency profile for each student (100 students were generated per condition) was computed based on an initial bias (e.g., "very low", "low", "any") and a sampling of the normal distribution across that bias. This approach does not yet account for students who may be more differentially prepared for the training (e.g., very low for some learning objectives, but high for others).

5 Results

We generated testbed simulations focused on the three questions introduced above. This section discusses a collection of tests, undertaken in the testbed, to explore each question.

Fig. 3 summarizes one analysis of potential learning gains for Question 1. It illustrates hypothesized learning curves for two different populations. The "medium" initial proficiency populations (dotted lines) are assumed to have some prior knowledge/familiarity of the domain, resulting in an overall higher level of initial proficiency for the EMT Scene Size-up unit. For example, such students might already be able to recognize certain visual cues in a given scene such as broken glass or fuel spills and be familiar with relevant categorization terms (*trauma victim*) relative to scene size-up. The other population is assumed to have very low initial proficiency (dashed lines), meaning that they have little relative working knowledge of the EMT domain.

The figure compares learning rates for a well-designed curriculum (purplish lines) to those obtained using targeted content selection (blue lines). In these examples, we assume tailoring to the learner is accurate and that content can be tailored to each learner (unlimited content options). These conditions provide a "best case" difference between a well designed CBT and an adaptive one. The results of the analysis suggest that the benefit from adaptive content selection is likely to be relatively modest in comparison to a well-designed, progressive CBT. We expected to see greater separation for the learners with low initial proficiency, but the relative gains between the two populations are similar. In general, these results suggest that a training effectiveness/pilot study for this domain will be highly sensitive to the initial instructional design. Either more tailoring opportunities or more learning time may be needed to better separate adaptive and non-adapted learner populations.

Fig. 4 summarizes exploration of trade offs between adaptive tailoring and the con-

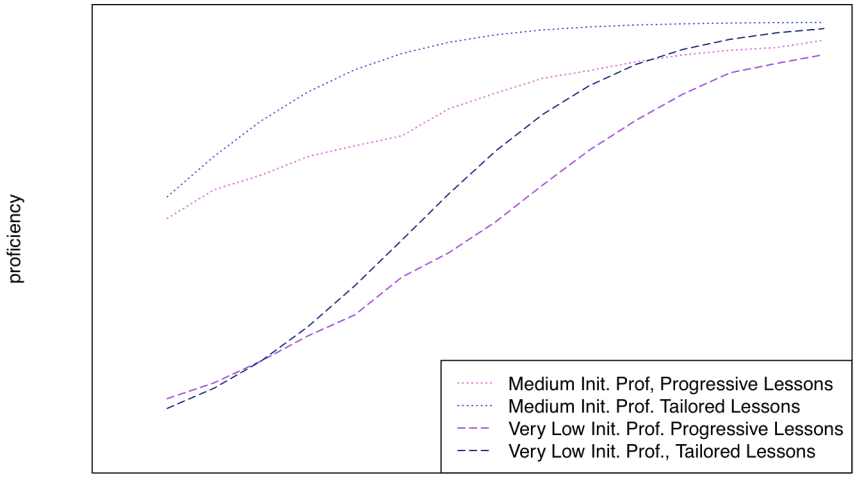


Fig. 3. Comparing Progressive (purple) & Tailored (blue) hypothesized learning trajectories for students with moderate a prior familiarity (dotted lines) and little familiarity (dashed).

tent available for adaptation. The figure contrasts projected learning outcomes under the same test conditions (other than available content) and uses the "very low" initial proficiency population as described for Fig. 3. The content options included in the figure are *unlimited* (content is available to match any proficiency profile) and a number of content choices: 2 choices (binary decision), 3-5 choices (small number of choices), and 10 choices (many choices). All choices were generated by sampling across the full spectrum of performance vectors. For example, for a 3 choice decision, one option would be generated for the "low", "medium", and "high" proficiency bias.

The figure suggests adaptive content selection is not likely to have a significant positive impact on learning unless sufficient content is available. Even 3-5 choices/decision were not sufficient to significantly improve learning. For continuing analysis, we plan to examine whether choices more localized to the typical learning progression (as reflected in the "progressive instructional design" in Fig. 3), could boost

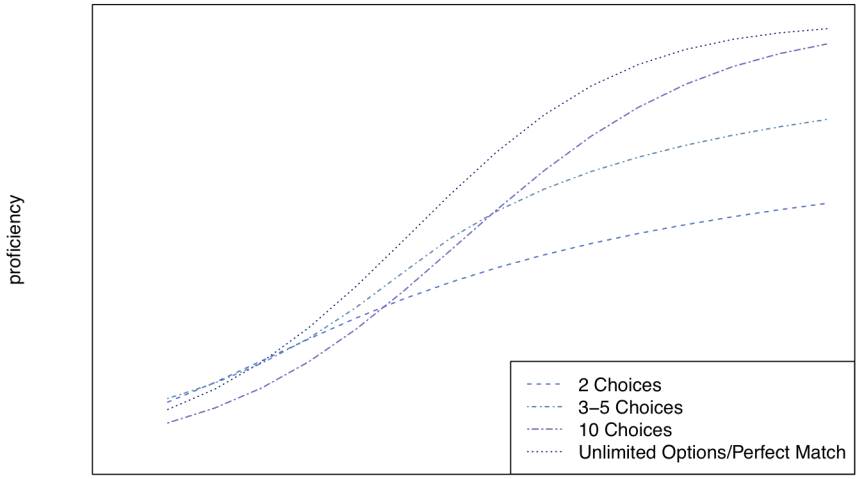


Fig. 5. The potential effects of content availability on learning outcomes.

the performance of adaptive content selection without requiring a prohibitive number of content options. In general, the worst-case performance for adaptive selection should be that of the original instructional design, so these results are somewhat more pessimistic than would be the case in actual implementation.

The final question was to attempt to quantify the accuracy of the underlying measures needed to enable adaptive tailoring. As shown in Fig. 2, we would like to use both explicit measures (e.g., a score from questions delivered after a lesson) as well as behavioral markers that provide (passive) indicators of learner state during learner activities in the CBT. Fig. 5 illustrates an initial assessment of the trade off inherent in using learner state measures to enable adaptive content selection. It presents learning curves obtained from a 95-70% range on measurement accuracy in

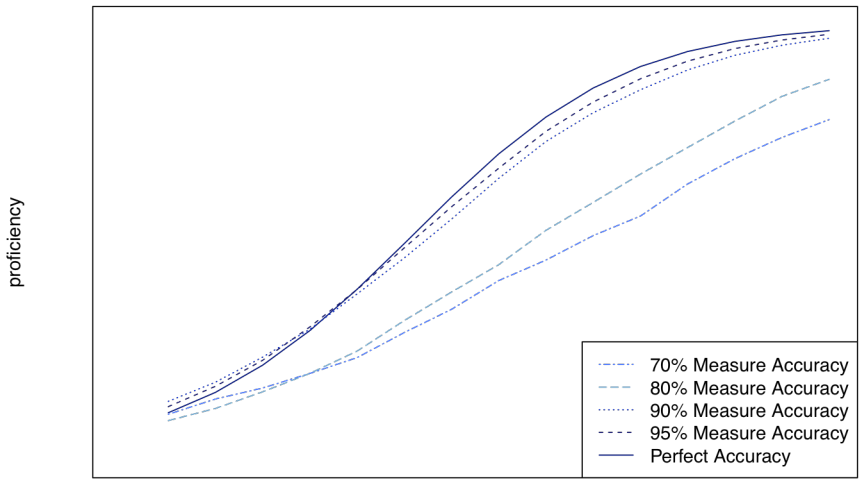


Fig. 4. The potential effects of measure accuracy on learning outcomes.

comparison to the learning curve obtained from perfect (100% accuracy) measures. Accuracy is computed as a normally distributed error around actual (ground-truth) levels of learner skill. It does not take into account compound errors across trials or reductions in measurement error with systematic, iterative measurement.

In general, as the accuracy of the measure degrades, the system's ability to narrow its tailoring to an individual learner's ZPD degrades as well. As suggested by the figure, even a (relatively good) 80% accuracy results in a loss of much of the advantage of adaptive content selection. This result, combined with the analysis summarized by Fig. 3, strongly suggests that adaptive content selection alone may not provide significant value for learning, given the limits of measurement accuracy, even if content requirement barriers could be mitigated (e.g., by some automatic content generation or content variation processes).

6 Conclusions

This paper illustrated an analytic approach to the design of adaptive training, enabling quantitative evaluation of design questions prior to commitments to implementation and pilot testing. In the illustrative example, analysis identified only marginal benefits of adaptive content selection in comparison to a well-designed learning environment. Further, realizing those small benefits requires unrealistic demands for accuracy in learner measurement and content creation. The results are somewhat discouraging from of the point of view of advancing adaptive training for this domain problem. However, more broadly, examples and tools supporting such analyses offer the potential to help researchers and practitioners set realistic expectations for learning system outcomes and to quantity component requirements within an adaptive training system to ensure minimum learning gains can be realized by an implemented system.

Acknowledgements

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs, through the Joint Program Committee-1/Medical Simulation and Information Science Research Program under Award No. W81XWH-16-1-0460. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense. The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office.

References

1. Landsberg, C.R., Astwood, R.S., Van Buskirk, W.L., Townsend, L.N., Steinhauser, N.B., Mercado, A.D.: Review of adaptive training system techniques. Military Psychology 24, 96-113 (2012)
2. Durlach, P.J., Lesgold, A.M. (eds.): Adaptive Technologies for Training and Education. Cambridge, New York (2012)

3. Folsom-Kovarik, J.T., Wray, R.E., Hamel, L.: Adaptive Assessment in an Instructor-Mediated System. Artificial Intelligence in Education (AIED), (2013)
4. Wray, R.E., Bachelor, B., Jones, R.M., Newton, C.: Bracketing human performance to support automation for workload reduction: A case study. LNCS: Proc. of HCII 2015 Conference. Springer-Verlag, Los Angeles (2015)
5. Wray, R.E., Woods, A., Haley, J., Folsom-Kovarik, J.T.: Evaluating Instructor Configurability for Adaptive Training Proceedings of the 7th International Conference on Applied Human Factors and Ergonomics (AHFE 2016) and the Affiliated Conferences. Springer, Orlando (2016)
6. Matsuda, N., Cohen, W.W., Sewall, J., Lacerda, G., Koedinger, K.R.: Predicting students' performance with SimStudent that learns cognitive skills from observation. In: Luckin, R., Koedinger, K.R., Greer, J. (eds.) Proc. International Conf. on Artificial Intelligence in Education, pp. 467-476. IOS Press, Amsterdam (2007)
7. MacLellan, C.J., Koedinger, K.R., Matsuda, N.: Authoring Tutors with SimStudent: An Evaluation of Efficiency and Model Quality. In: Proceedings of the 12th International Conference on Intelligent Tutoring Systems, pp. 551-560. (Year)
8. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S.A., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111, 1036-1060 (2004)
9. Newell, A., Rosenblum, P.S.: Mechanisms of skill acquisition and the law of practice. In: Anderson, J.R. (ed.) Cognitive Skills and Acquistion. Erlbaum, (1980)
10. Eagle, M., Corbett, A., Stamper, J., McLaren, B.M., Wagner, A., MacLaren, B., Mitchell, A.: Estimating individual differences for student modeling in intelligent tutors from reading and pretest data. International Conference on Intelligent Tutoring Systems, pp. 133-143. Springer (2016)
11. Lee, J.I., Brunskill, E.: The Impact on Individualizing Student Models on Necessary Practice Opportunities. Intrntl Educational Data Mining Society, (2012)
12. Durlach, P.J., Ray, J.M.: Designing Adaptive Instructional Environments: Insights from Empirical Evidence. Army Research Institute for the Behavioral and Social Sciences (2011)
13. Schatz, S., Oakes, C., Folsom-Kovarik, J.T., Dolletski-Lazar, R.: ITS + SBT: A Review of Operational Situated Tutors. Military Psychology, special issue on current trends in adaptive training for military ap-plication (2012)
14. Lane, H.C., Johnson, W.L.: Intelligent Tutoring and Pedagogical Experience Manipulation in Virtual Learning Environments. In: Cohn, J., Nicholson, D., Schmorrow, D. (eds.) The PSI Handbook of Virtual Environments for Training and Education, vol. 3. Praeger Security International, Westport, CT (2008)
15. Chaplot, D.S., Rhim, E., Kim, J.: Personalized Adaptive Learning Using Neural Networks. In: Proc. 3rd ACM Conference on Learning @ Scale. (2016)
16. United States Department of Transportation, National Highway Traffic Safety Administration: EMT-Basic: National Standard Curriculum. (1996)
17. Leibowitz, N., Baum, B., Enden, G., Karniel, A.: The exponential learning equation as a function of successful trials results in sigmoid performance. Journal of Mathematical Psychology 54, 338-340 (2010)
18. Murray, T., Arroyo, I.: Toward Measuring and Maintaining the Zone of Proximal Development in Adaptive Instructional Systems. Proceedings of the 6th International Conference on Intelligent Tutoring Systems, pp. 749-758. Springer-Verlag (2002)

APPENDIX B: PROTOCOL FOR HUMAN SUBJECTS STUDY

Assessing the Role of Behavioral Markers in Adaptive Learning

Study Protocol
Version 1.5.8 (8 May 2017)

Robert E. Wray, PhD (PI)
Soar Technology, Inc.

A. Introduction and Background

Personalized learning, in which a learning environment adapts to the abilities, needs, and preferences of individual learners, has been identified as a "Grand Challenge" for 21st century research and engineering (National Academy of Engineering, 2008). The benefits of adaptive learning environments include more efficient learning (Woolf, 2008), improved attention and motivation (Craig et al., 2004), the development of less rigid and more flexible decision making (i.e., adaptive expertise, Hatano & Inagaki, 1986), and improved transfer of learning to settings in which learned knowledge is used and applied (Bransford & Schwartz, 1999; Coultas, Grossman, & Salas, 2012; Pan & Yang, 2010). Improved and personalized learning has particular application for more pervasive and less costly medical training, which often is delivered primarily by human instructors in classes with modest student-to-teacher ratios. Human instruction and mentoring is very valuable and desirable, but adaptive personalization methods offer an opportunity to deliver good, effective introductory and basic training, thus potentially enabling a single human instructor to train many more students by better preparing them for coaching and instruction from experts.

Adaptation to a learner usually requires a model of the learner that is frequently updated as a learner progresses through a curriculum (Durlach & Spain, 2012). The targeting of adaptive techniques, such as scaffolding (Pea, 2004) and competency matching (Murray & Arroyo, 2002; Vygotsky, 1978), depends on the accuracy (and, to some degree, precision) of the learner model. When the model better reflects the learner's actual knowledge, skills, and attitudes at any point during the learning, the targeting of the adaptive method to the learner generally improves (Murray & Arroyo, 2002).

Creating a complete and accurate learner model is difficult, however. In addition to estimating learner capability from formal and informal assessment within the environment (Anderson et al., 1995; Dillenbourg & Self, 1992; Durlach & Spain, 2012; Pardos et al., 2010), researchers have explored many behavioral, physiological, and even neurological indicators or "markers" that can provide additional context for estimating a learner's cognitive state and improving the dynamic assessment of the learner . For example, behavioral sensors (posture, eye trackers), physiological sensors (Galvanic skin response), and neurological sensors (EEG) have all been used to assess and track learner arousal/attention in learning environments (Cohn, Nicholson, & Schmorrow, 2008). Further, understanding the dynamic patterns of learner attention/arousal allows the identification of dynamic adaptation targeted to the identified arousal states (Cohn, Kruse, & Stripling, 2005).

Such markers can be useful for improving a learner model, but most markers today require sensors that are not commonly available on the hardware available for typical computer-based learning: a laptop or a tablet. The primary goal of this study is to assess the role of behavioral markers that have the potential to improve learner modeling while also not requiring specialized hardware/sensors (i.e., using only hardware sensors found on typical computing devices). The study focuses specifically on behavioral markers that can be derived from 1) mouse movements and mouse selections (“clicks”) and 2) patterns of eye movements observable from a web camera (“passive eye tracking”).

There is significant and growing scientific evidence that the temporal patterns of mouse movements during selection tasks can provide reliable insight into the cognitive state of subjects (Hehman, Stolier, & Freeman, 2015; Quétard et al., 2016). We anticipate, however, these markers to be noisier (less diagnostically precise) than neuro-cognitive markers associated with specialized sensors. Thus, this study focuses on evaluating the impact of the behavioral markers on the adaptive learning system to improve learning outcomes, given the noise and uncertainty of measure inherent in these unspecialized sources.

Under this study, multiple hypotheses will be explored:

- H1: There is a difference between conditions such that learning outcomes from the adaptive condition will exceed those from the non-adaptive condition.
- H2: Mouse movements will be an indicator of learner focus on certain aspects of the learning environment.
- H3: Eye movements will be an indicator of learner focus on certain aspects of the learning environment.
- H4: Mouse and eye movements will be correlated.

The proposed study is being funded by the United States Army Medical Research Acquisition Activity under the title *Applied Cognitive Models of Behavior and Errors Patterns* (Grant number W81XWH-16-1-0460).

B. Study Design

In order to explore the hypotheses discussed in section A, a research study will be implemented which compares the results of learning between an adaptive medical learning unit to a unit presented in a non-adaptive (fixed) sequence. Specifically, curriculum units will be developed for "Scene Size Up," a required curriculum component used in Emergency Medical Technician (EMT) training (United States Department of Transportation & National Highway Traffic Safety Administration, 1996). These units (both adaptive and non-adaptive) will be presented to university subject population(s) in order to assess the utility of markers to improve adaptive learning in emergency medical environments. As discussed in section E, we will use multiple routes of recruitment, which will allow us to complete the study between July 1st, 2017 and January 31st, 2018.

Specifically, the following variables of interest will be implemented and observed:

- **Instructional approach**: The overall instructional approach of the learning environment. For this study, there are two distinct instructional approaches:
    - **Non-adaptive/traditional**: An instructional unit that is presented in a fixed sequence to all learners.
    - **Adaptive based on performance (only)**: An instructional unit in which specific content presentations are constructed/chosen based on learner performance and subsequent estimates of learner knowledge and skill.
    - **Adaptive based on performance and markers**: An instructional unit that is dynamically constructed/chosen based on a combination of direct learner observation (as above) and behavior markers.
- **Markers**: Patterns of observed behavior that are hypothesized to have a role in improving a learner model.
- **Knowledge gain**: A measure of the post-test performance of subjects, relative to pre-test performance.

This study will be implemented as a between-subjects design, with "instructional approach" being the independent variable of interest. Instructional approach will be manipulated at three levels (as discussed above): non-adaptive, adaptive based on performance (only) and adaptive based on performance and markers. To maintain the integrity of results, assignment will be randomized, with neither participants nor the experimenter being aware of assignment ahead of time.

Primary Experimental Conditions

Non-Adaptive (Standard Presentation)

Adaptation (Performance)

Adaptation (Performance and Markers)

The primary dependent variable will be "knowledge gain", as measured by difference scores between pre- and post-tests given to participants. Additionally, the behavioral markers outlined in section A, derived from dynamic tracking of mouse movements and eye movements, will be used to predict learner needs and adapt the learning environment. The combination of these variables will enable the study to address the hypotheses above, as well as quantify the utility of the chosen adaptive learning models for improving learning in medical environments.

C. Procedure

The procedure implemented for participants in this study is expected to take between 45 and 75 minutes. Specific steps in the procedure are detailed chronologically below.

1. Upon arrival, participants will read and sign the informed consent document.
2. Once participants have indicated their consent, they will be randomly assigned one of the three experimental conditions.
3. All participants will be given a standard demographics questionnaire (Appendix A) to assess their education level and familiarity (if any) with EMT training or medicine.
4. All participants will receive a short 5-minute tutorial on how to use APACTS (see Appendix B).
5. Passive eye tracking and mouse tracking mechanisms will be calibrated during the tutorial. Calibration includes the following standard practices:
    1. For eye tracking, adjustment of cameras and gaze calibration will be completed. This will require minimal activity from the participant, such as being asked to look around the screen (see Appendix B for example).
    2. For mouse tracking, calibration of the mouse will be completed. This will require minimal activity from the participant, such as being asked to move the mouse around the screen (see Appendix B for example).
6. All participants will complete a pre-test, developed by the experimenters, which contains questions about the process of completing the scene size-up task as an EMT (see Appendix C).
7. In their assigned condition, participants will learn how to complete a scene size-up, which will include the following standard practices for EMT training (see Appendix D for example content):
    1. Learning scene size-up terms and associated tasks.
    2. Viewing images of emergency scenes and reading text-based descriptions of the emergency scenes viewed.
    3. Viewing images of emergency scenes with opportunities to practice concepts learned, such as answering a question or labeling areas in a displayed image.
8. During their completion of these conditions, passive eye tracking and mouse tracking will be engaged to collect participant data.
    1. In the adaptive conditions, results from passive eye tracking and mouse tracking will be used to change what content is presented to the learner, such as varying the difficulty of practice tasks, presenting feedback customized to a subject's response, and/or repeating or amplifying previously presented information.

    2. In the non-adaptive condition, the content presentation will not differ; all subjects will receive the same information, with identical feedback and level of difficulty as all other subjects.
9. During completion of conditions, participants will also receive questions tracking their sense of progress / self-efficacy in the domain.
10. Participants will complete a post-test, which will be identical to the pre-test (Appendix C).
11. Participants will be given an opportunity to give verbal feedback about the study before they leave.

D. Inclusions / Exclusion Criteria

The following inclusion/exclusion criterion will be adhered to and verified for each participant:

- Must be 18+ years old

The primary population of subjects will be college students, due to the source of recruitment (detailed in section E). College students represent a apt population for studying professional (in this case EMT) training, as they are pursuing professional endeavors that require similar training and learning practices. At the same time, the principle of distributive justice applies in this context, as college students represent a low risk population that can benefit from participation in research (through class credit or payment; see section E), and the study research is likewise low risk.

E. Recruitment of Participants

Primary Study Site: University of Alabama

The primary source of participants is the University of Alabama. Participants will be recruited from the University of Alabama through 3 different methods:

- Volunteers from University of Alabama's GBA300 classes, who are able to receive class credit for participation.
- Volunteers from University of Alabama's research participant pools, including Psychology Sona and CCIS participant pool, which are used to grant class credits.
- Paid participants recruited through flyers posted through University of Alabama's campus and on social media websites (see Appendix E).

Recruitment will begin in August 2017, with flyers/announcements being posted in classes and listed in the participant pools (per above list). We will not be requesting a set number of participants from each source. Instead, participants will be recruited freely through the above methods until the required sample size is met (see section I). Recruitment will be performed by the sub-investigator on the project, who has CITI certification through completing the "Group 2: Social Behavioral and Education Research Investigators and Key Personnel" course.

Secondary Study Site: Soar Technology, Inc. (Orlando Office)

Some subjects, especially for initial system testing and pilot assessment, will be recruited from the University of Central Florida (UCF) and Research Park areas. These subjects will exclusively be paid participants recruited through flyers posted through UCF’s campus, Research Park (adjacent to UCF), as

well as email and social media websites (see flyer in Appendix E). Recruitment will be coordinated by both the Principal Investigator (Wray) and the sub-investigator (Stowers). Both have CITI certification. Subjects recruited at UCF will complete the study at the Orlando offices of Soar Technology, which is located in Research Park. An office will be dedicated for data collection at Soar Technology.

F. Consent Process and Timing

Consent will be obtained upon participant arrival to the research site. Before beginning the study, participants will be given a copy of the informed consent to read (the consent form will be developed by E&I for this study and thus is not attached to this submission). The experimenter will also explain the consent to them verbally. Participants will be given as much time as they need to consider participation and will consent verbally, as well as through written signature, before proceeding with the study.

The consent process will be performed the PI, the sub investigator and research assistants. All experimenters will have CITI "Group 2: Social Behavioral and Education Research Investigators and Key Personnel" certification.

G. Risks, Discomforts, and Benefits to Subjects

Minimization of Risks

Due to the nature of content used in the study, participants may find some of the images in the study disturbing (accident victims). These risks will be minimized through the use of images that minimize the visible presentation of injuries.

Maximization of Benefits

Participants will learn how to assess a medical emergency, and may find that learning process intrinsically rewarding. Benefits will be maximized through the use of practice rounds, as well as pre-tests and post-tests, where participants will be able to demonstrate their success in learning the content presented.

Provisions to protect the privacy of participants:

Privacy of Participants and Confidentiality of Data

Participant information will only be identified through assigned identification numbers. Through the use of the identification numbers, the data will be fully anonymous. Information connecting identification numbers with any personally identifiable information will be held in a separate location from other data collected and stored on a password protected computer. Only those involved in the study will have access to any information or data linked to the study.

Data Storage

Data will be stored for 5 years, according to guidelines by CITI. Data will be stored on a password-protected computer at all times and only the principal investigator and sub-investigator will have access to individual data.

H. Financial Considerations

Participants will be compensated $15 for participation via a credit-card gift card. Compensation will be provided at the end of the experimental session. Participants are not expected to incur any costs to themselves as a result of participation. If any research related injuries are discovered, the principal investigator and IRB will be notified immediately, as well as the University of Alabama's counseling and medical centers. Participants will have direct access to health care and counseling as needed.

I. Data Analysis and Statistical Analysis

As this study involves a single independent variable with just three levels, the primary analysis will be an F test comparing the difference scores of pre- and post-tests in each condition. Additionally, correlations will be calculated in order to gain an understanding of the relationship between behavioral markers and performance outcomes. A power analysis was run (using GPower 3.1) based on the following criteria:

- F test (one-way ANOVA)
- Effect size (*f*): 0.4
- Error probability (*alpha*): 0.05
- Power (*1 - beta error probability*): 0.85
- Number of groups: 3

According to the parameters entered and calculations made using GPower, we will need to analyze data from 72 participants to achieve optimal power. In order to account for participant withdrawal, as well as any issues encountered with eye tracking or mouse tracking that may cause data to be unusable (e.g., an adaptive condition in which mouse tracking did not function), we will collect data from up to 100 participants.

Analyses of participant data will be broken up into the following steps, the final step marking the endpoint of the study:

1. Coding and cleaning mouse-tracking and eye-tracking data
2. Calculating difference scores for pre- and post-tests
3. Calculating t-test and correlations
4. Reporting results through technical reports and publications

Our expectation is that all primary data analysis will be concluded by April 30, 2018. However, as data will be kept up to 5 years past the end of collection (see section G), we expect to also analyze depersonalized data on an ongoing basis. In particular, we will data captured from eye tracking and mouse tracking to inform further development and refinement of the markers tested in this study. For example, we are focusing a single mouse-tracking algorithm for use in the study. After the study is completed, we can perform post-hoc analysis with participant mouse tracking data to evaluate alternative mouse tracking algorithms and possible pattern-based selection of algorithms for future studies. Thus, the data resulting from this experiment will support subsequent research and improvement of adaptive learning methods and tools.

References

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

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking Transfer: A Simple Proposal With Multiple Implications. In A. Iran-Nejad & P. D. Pearson. (Eds.), *Review of Research in Education* (Vol. 24). American Educational Research Association: Washington, DC.

Cohn, J. V., Kruse, A., & Stripling, R. (2005). Investigating the transition from Novice to expert in a virtual training environment using neuro-cognitive measures. In D. Schmorrow (Ed.), *Foundations of Augmented Cognition*. Mattawan, NJ: LEA.

Cohn, J. V., Nicholson, D., & Schmorrow, D. (Eds.). (2008). *The PSI Handbook of Virtual Environments for Training and Education* (Vol. 3). Westport, CT: Praeger Security International.

Coultas, C. W., Grossman, R., & Salas, E. (2012). Design, Delivery, Evaluation, and Transfer of Training Systems *Handbook of Human Factors and Ergonomics*: John Wiley & Sons, Inc.

Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media, 29*(3), 241-250.

Dillenbourg, P., & Self, J. (1992). A framework for learner modeling. *Interactive Learning Environments, 2*(2), 111-137.

Durlach, P., & Spain, R. (2012). Framework for Instructional Technology. In V. G. Duffy (Ed.), *Advances in Applied Human Modeling and Simulation*: CRC Press.

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Asuma & K. Hakauta (Eds.), *Child Development and Education in Japan* (pp. 262-272). San Francisco: Freeman.

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes and Intergroup Relations, 18*(3), 384-401.

Murray, T., & Arroyo, I. (2002). *Toward Measuring and Maintaining the Zone of Proximal Development in Adaptive Instructional Systems*. Paper presented at the Proceedings of the 6th International Conference on Intelligent Tutoring Systems.

National Academy of Engineering. (2008). *Grand Challenges for Engineering*. Washington, DC: National Academy of Sciences/National Academy of Engineering.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345-1359.

Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan, C. L. (2010). Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks. In R. Christobal (Ed.), *Handbook of Educational Data Mining* (pp. 417-426): CRC Press.

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences, 13*(3), 423-451.

Quétard, B., Quinton, J. C., Mermillod, M., Barca, L., Pezzulo, G., Colomb, M., et al. (2016). Differential effects of visual uncertainty and contextual guidance on perceptual decisions: Evidence from eye and mouse tracking in visual search. *Journal of Vision, 16*(11).

United States Department of Transportation, & National Highway Traffic Safety Administration. (1996). EMT-Basic: National Standard Curriculum.

Vygotsky, L. S. (1978). *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Woolf, B. P. (2008). *Building intelligent interactive tutors: student-centered strategies for revolutionizing e-learning*: Morgan Kaufman.

Appendix A

Demographics Questionnaire

1. How old are you?

- __ (Fill in the blank)

2. Are you male or female?

- male
- female
- other

3. What is your education level?

- Graduated high school
- Completed some college coursework
- Completed Associate's degree
- Completed Bachelor's degree
- Completed Master's degree
- Completed Doctoral degree
- Other (please explain)
    - __ (fill in blank)

4. What is your major of study?

- __ (Fill in the blank)

5. Do you have any training or experience as an emergency medical technician or related service?

- Yes
- No

6. Do you have any formal training in first-aid procedures (such as a CPR course or training as a lifeguard)?

- Yes
- No

7. If yes to Question 5 or 6, please sketch some details (what training, when, etc.).

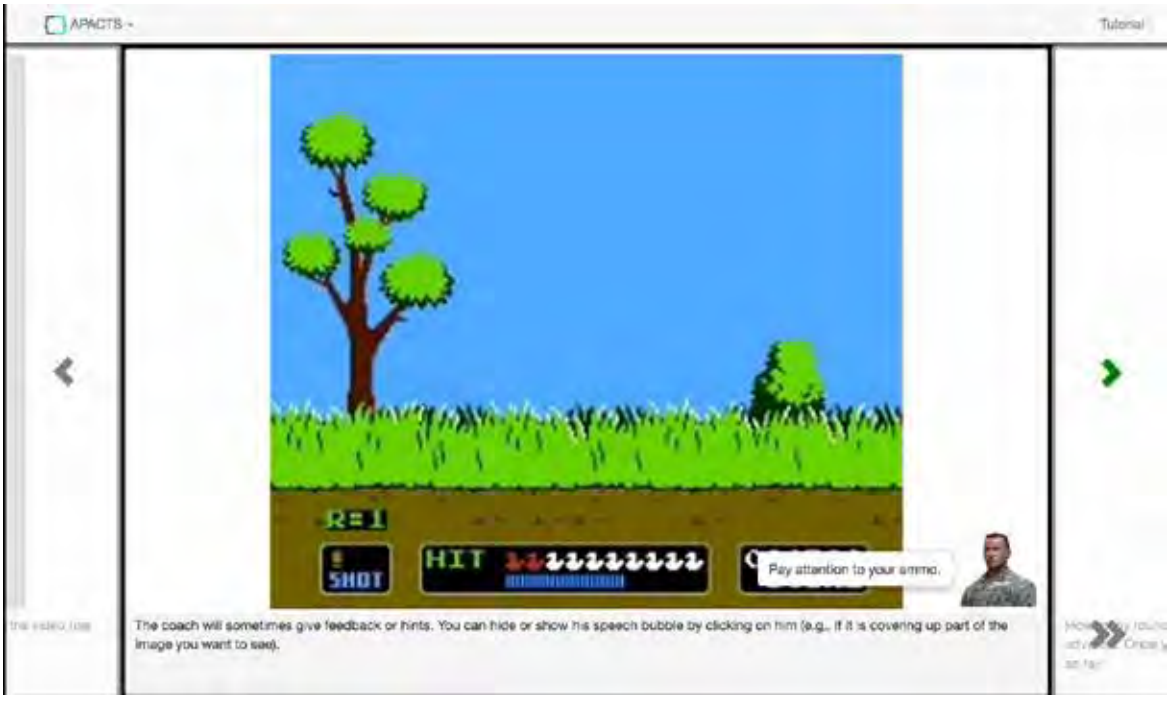
- __ (Fill in the blank)

Appendix B

Environment Tutorial & Calibration

Standard tutorial introduction to the instructional content delivery system (APACTS)

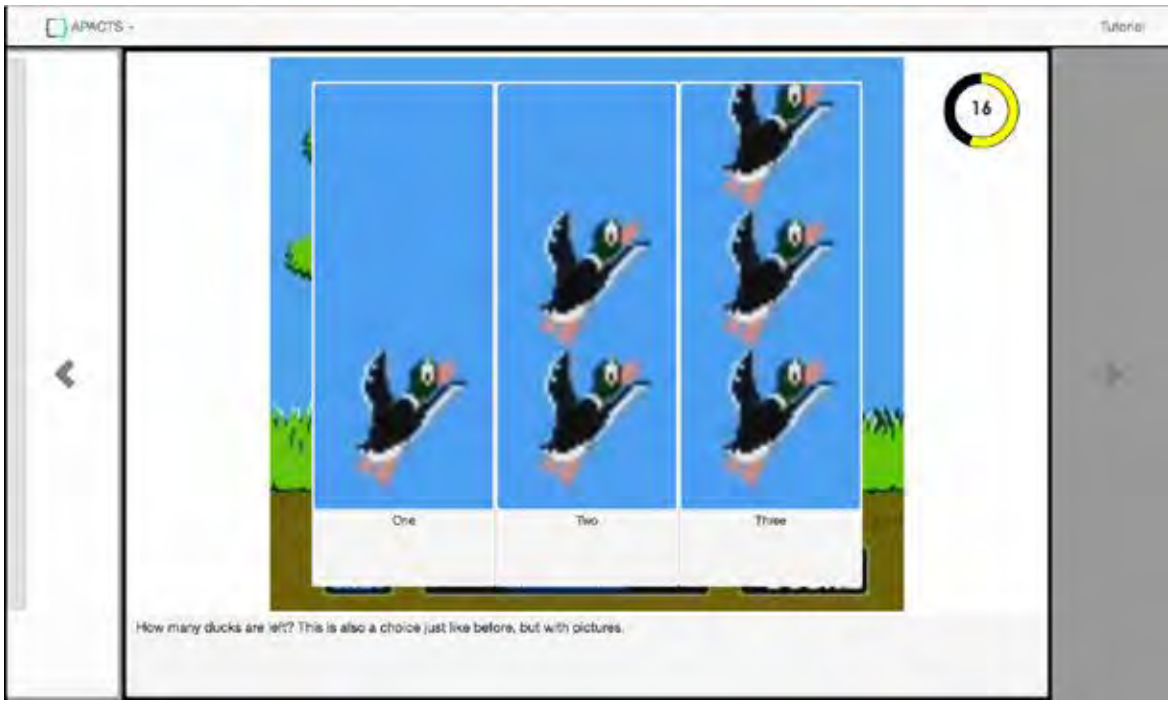
APACTS supports embedded videos



The “Coach” is used to provide directions, amplifying information, additional explanation, etc.

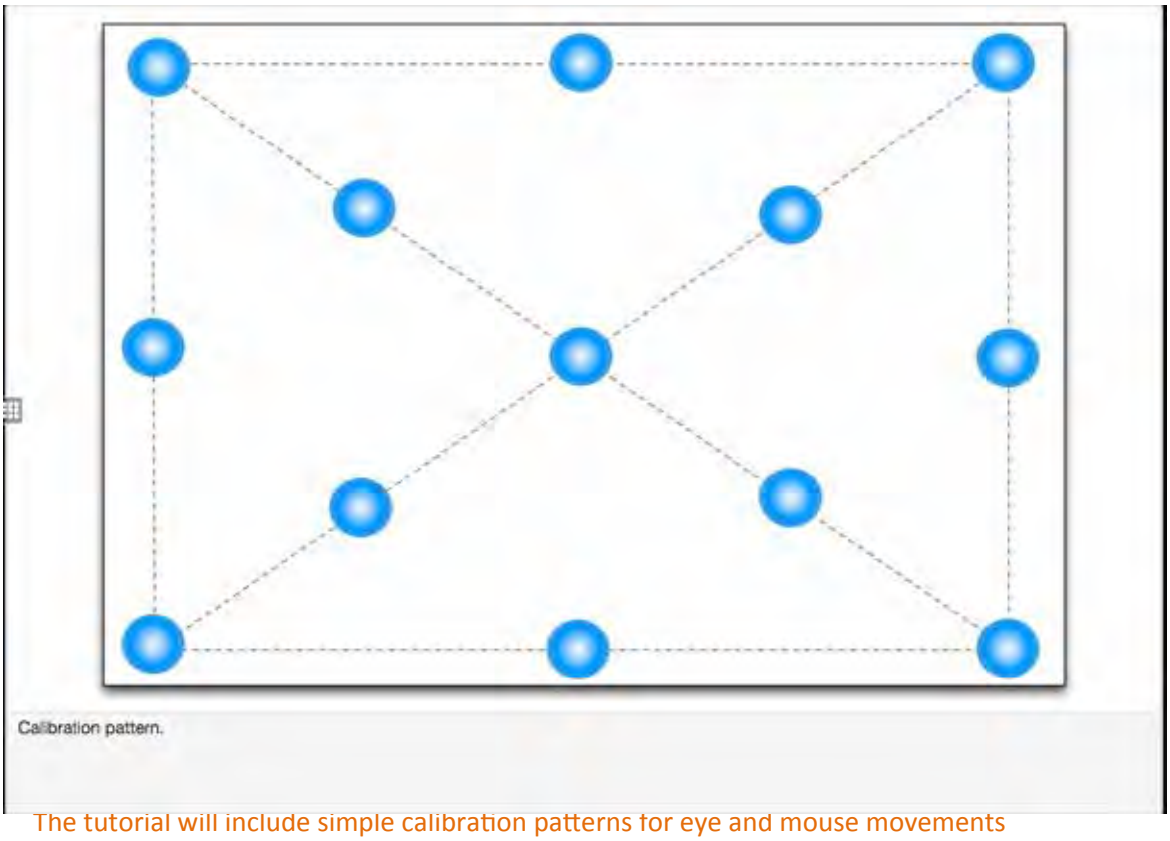
Introducing a “choice frame” (multiple choice questions)

Introducing annotation frames (tag locations within an image)



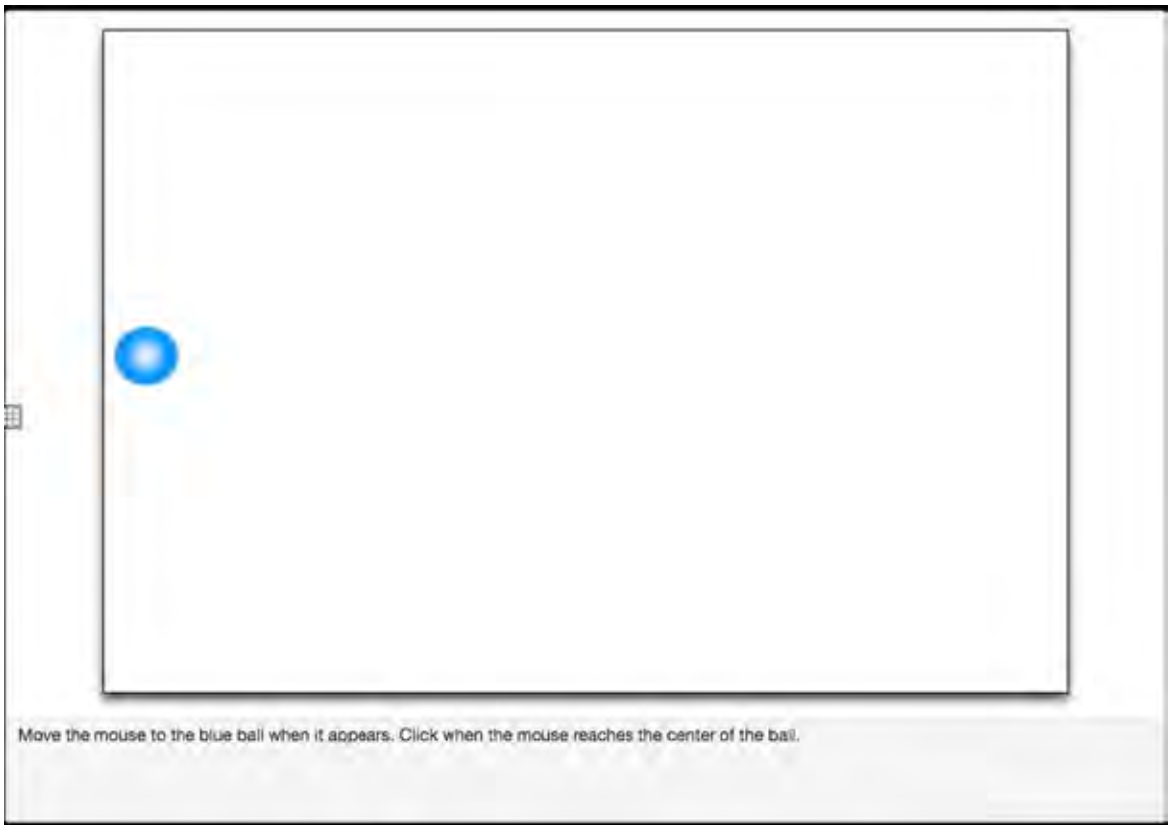
Choice frames can include images and text.

An alternative annotation frame



The tutorial will include simple calibration patterns for eye and mouse movements
(This image shows the underlying calibration pattern.)

The actual calibration task will be 1) to fixate on a series of screen locations based on pattern, …



And 2) to move the mouse to a subsequence series of screen locations.

Appendix C. Pre-Test/Post-Test Example Questions

Subjects will complete a pre-test and post-test as part of the study. The pre-test and post-test will both be administered within the computer-based learning environment in which learning content is delivered (see Appendix D for specific examples of how questions are delivered within the system).

The pre-test and post-test will be identical and will not include any adaptive choices (the specific questions and their order will be fixed for all subjects/experimental conditions).

Below, we provide examples of the pre-/post-test questions for the study.

Basic Conceptual Knowledge

1. Which of the following best expresses the definition of *mechanism of injury* (MOI)?
   (a) The types of injuries observed for particular kinds of accidents
   (b) The immediate cause(s) of an injury that results from an accident
   (c) Mechanical failures in a vehicle (e.g., a blow out) that result in accident and injury
   (d) Action(s) that lead to accident and injury (failure to yield)
   (e) Both (c) and (d)

2. Which option best describes when scene size-up should be undertaken?
   (a) As soon as possible after arrival, but after immediate patient triage
   (b) During transit to the accident location, as provided by emergency personnel on scene via radio (or similar)
   (c) Immediately on arrival
   (d) After hazards have been assessed and bystanders moved away from hazards
   (e) Both (a) and (d)

3. What patterns of injuries are associated with side-impact collisions?
   (a) Head and neck injuries
   (b) Knee, hip, and leg injuries
   (c) Direct, blunt trauma
   (d) Broken arms and ribs
   (e) Both (a) and (b)
   (f) Both (a) and (c)
   (g) (a), (b) and (c)

4. What pattern(s) of injury are most associated with the "Down and Under" mechanism of injury?
   (a) Head and neck injuries
   (b) Knee, hip, and leg injuries
   (c) Direct, blunt trauma
   (d) Broken arms and ribs
   (e) Both (a) and (b)
   (f) Both (a) and (c)

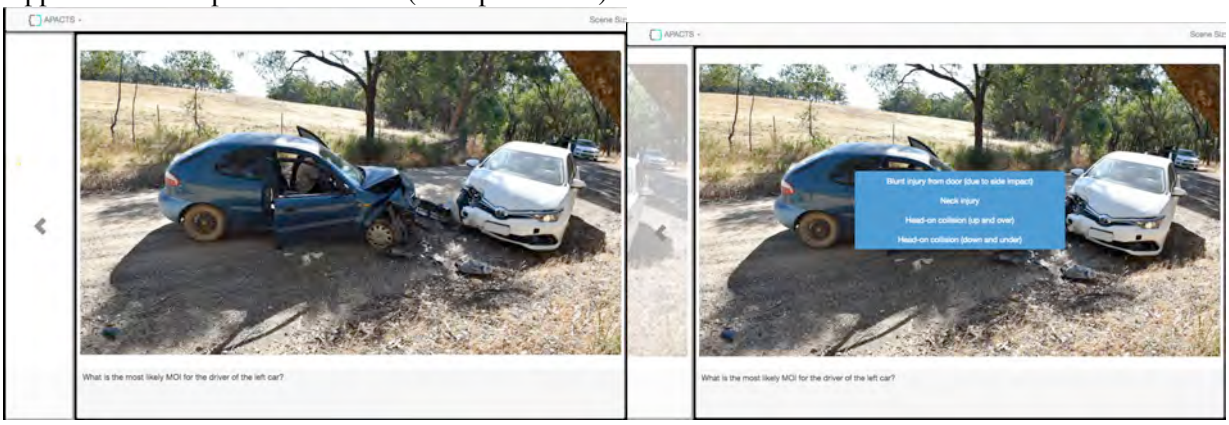
4. What pattern(s) of injury are most associated with a roll over mechanism of injury?
   (a) Head and neck injuries
   (b) Knee, hip, and leg injuries
   (c) Direct, blunt trauma

(d) Broken arms and ribs
(e) Both (a) and (b)
(f) Both (a) and (c)
(g) All of the above

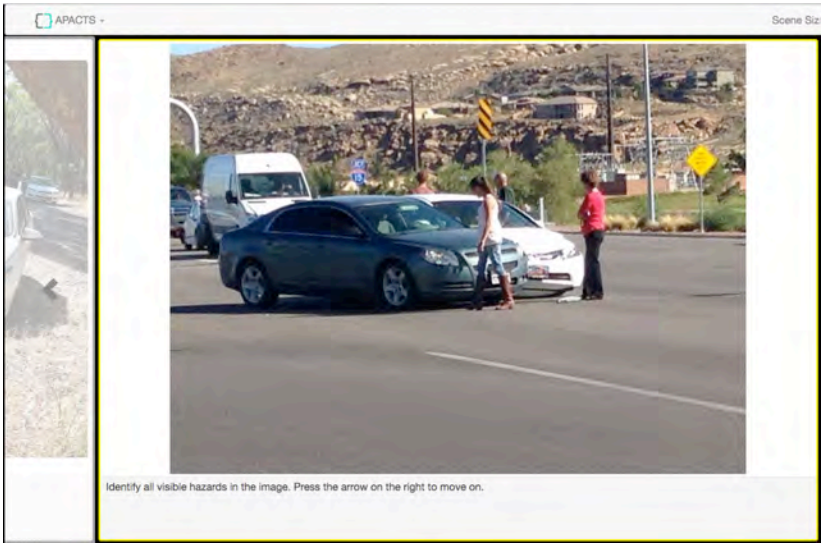
In addition to general knowledge questions, the pre- and post-test will include questions that present an image of an accident and ask the subject to evaluate the situation (size up the scene) in accordance with materials presented in the learning unit. These questions will be similar to the assessment and feedback questions that are used within the learning environment (i.e., as summarized in Appendix D).

Examples:

Application to a specific situation (multiple choice)

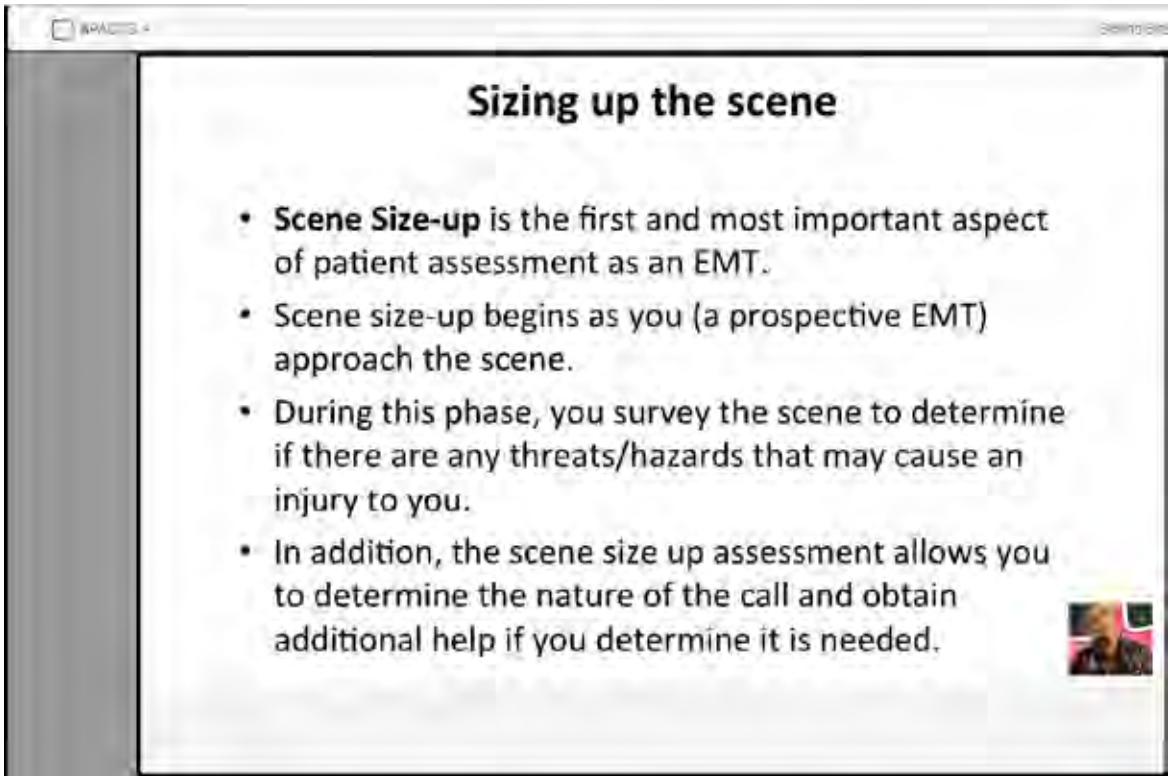


Application to a specific situation (labeling/annotation)

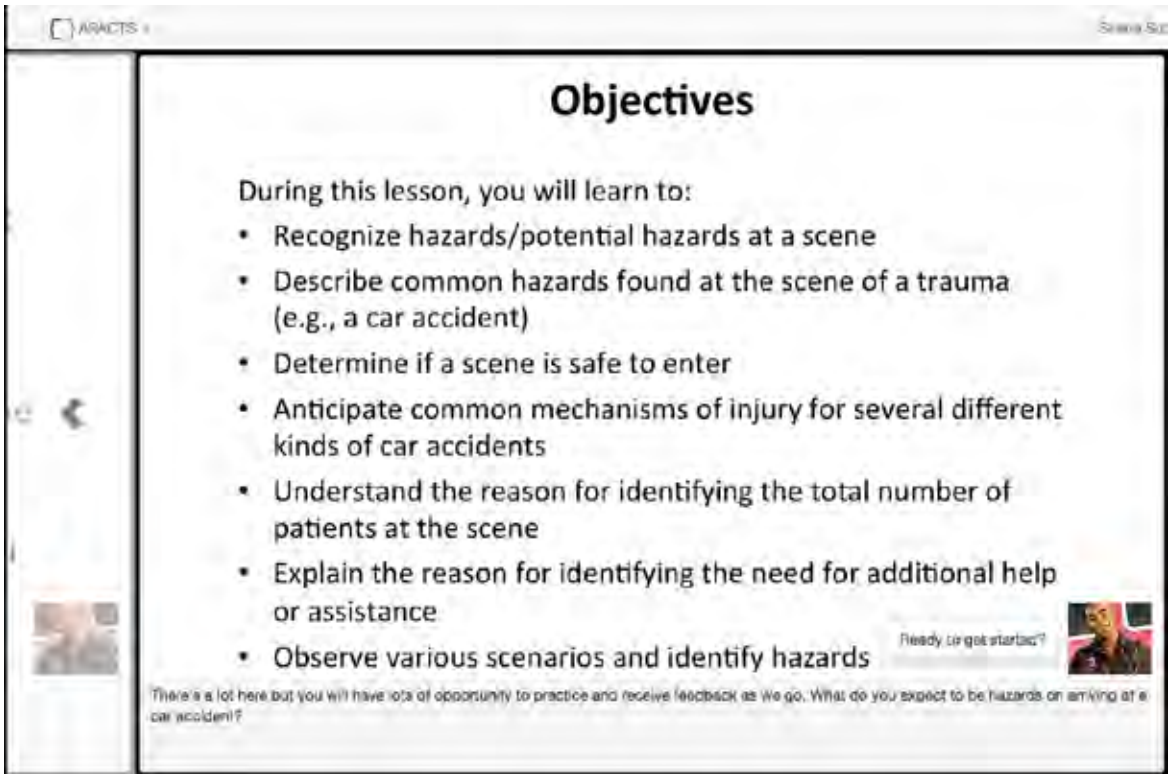


Appendix D

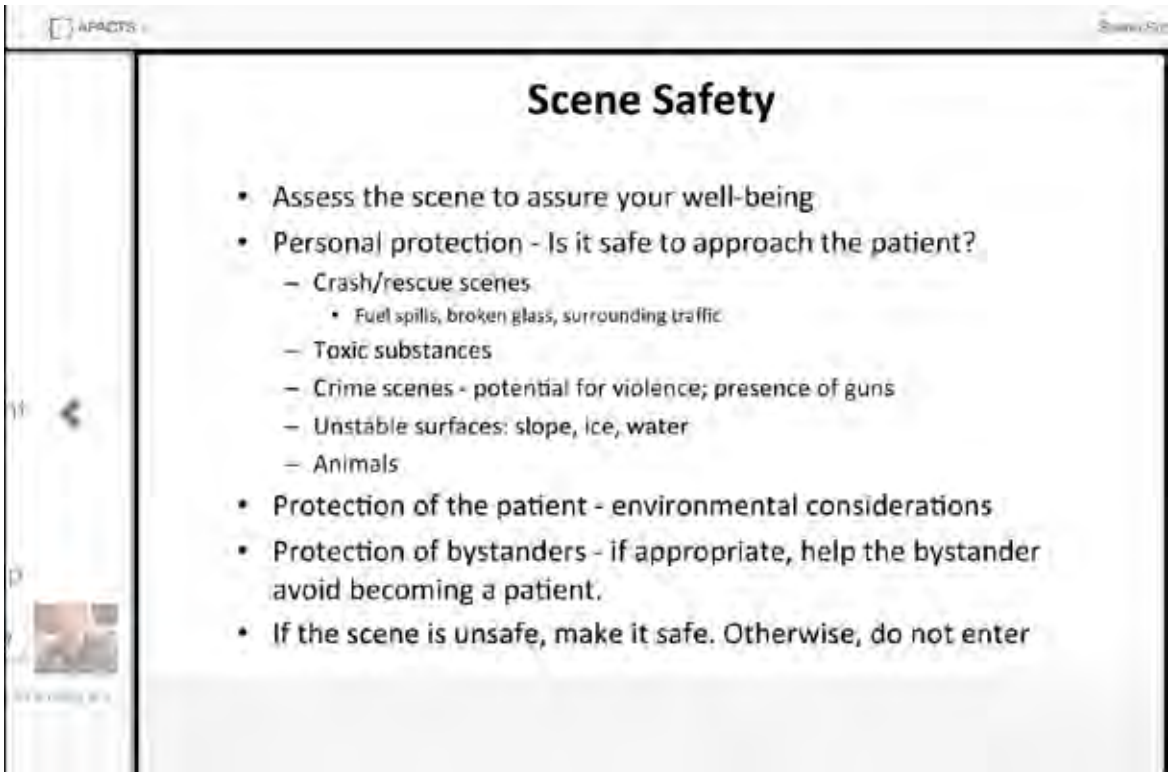
Example Content from the
Learning Environment



Introductory instructional material

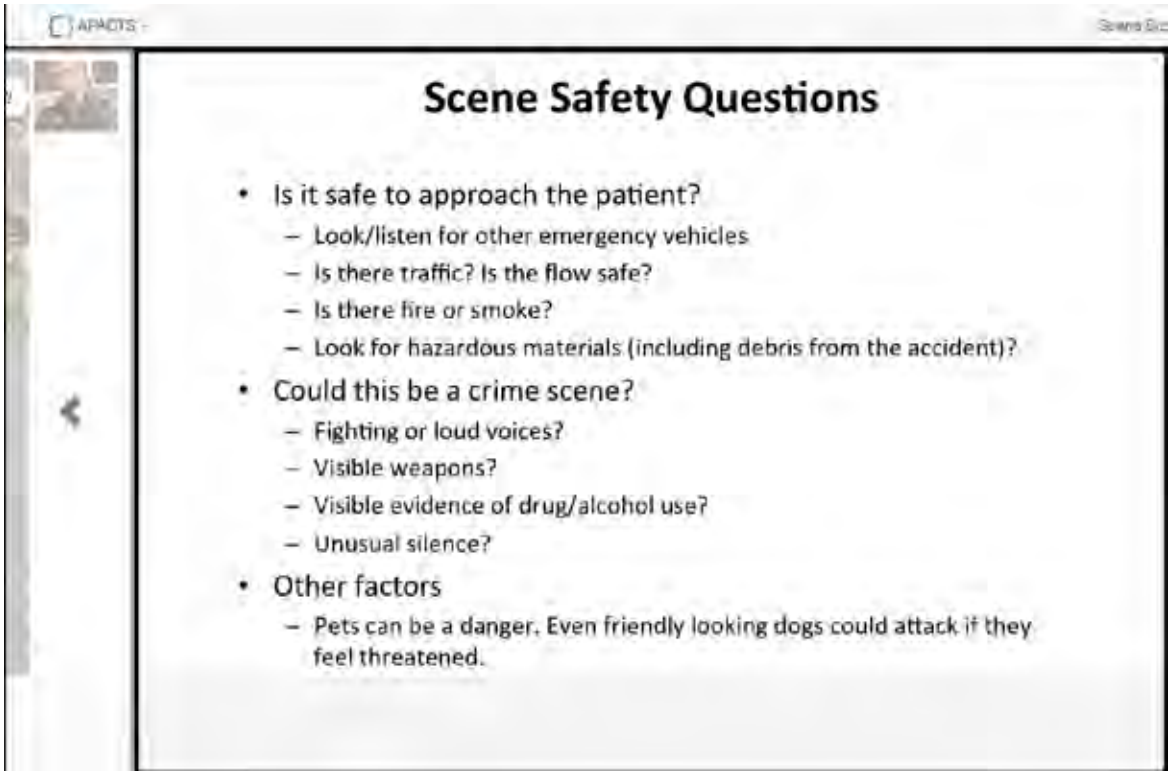


Objectives of the unit of study. Clicking on the "coach" will bring up amplifying or summary statements.



More detailed lesson material.

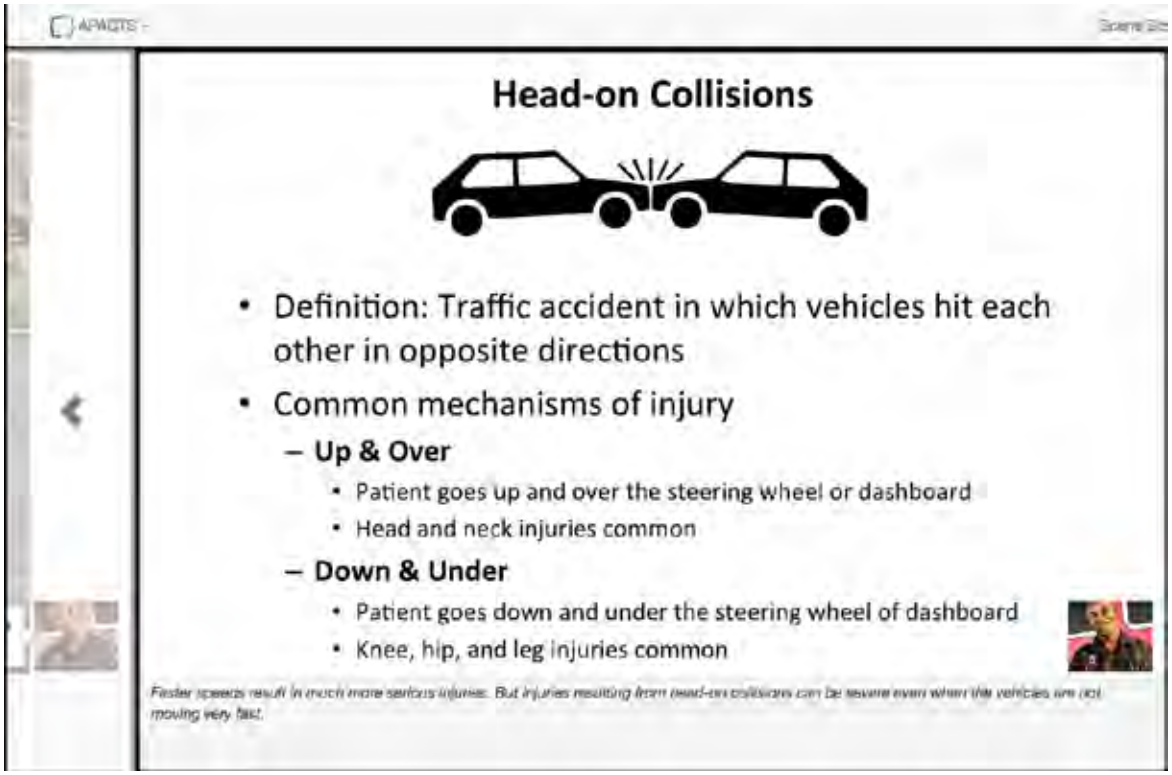
Opportunity to anticipate and consider more detailed explanation.



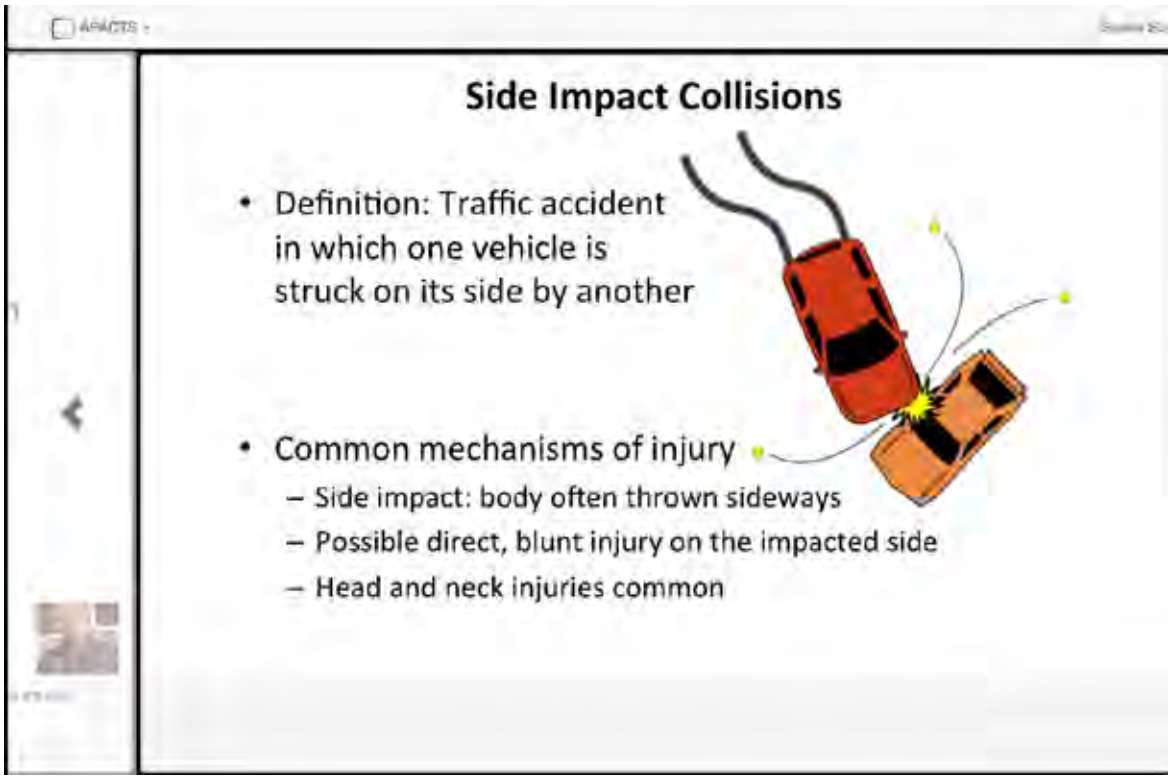
"Check your knowledge" questions. Responses to these questions are used to update the learner model and influence subsequent content choices.

Simulated user response…

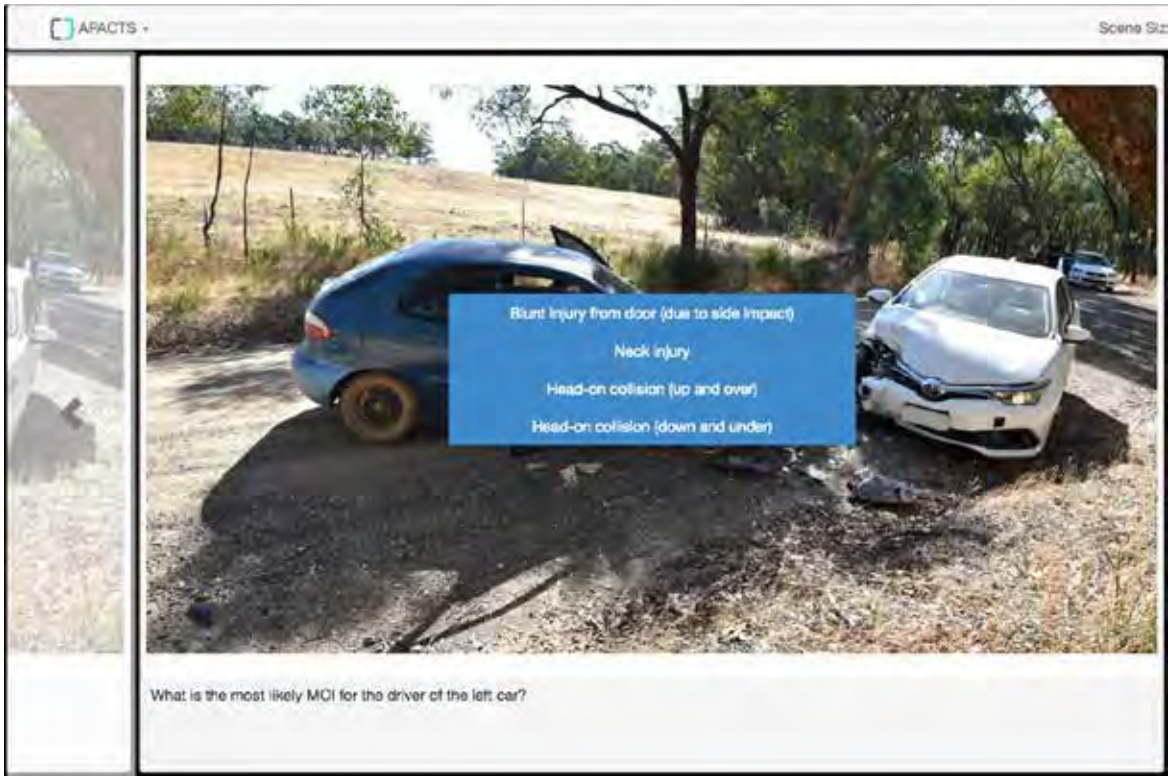
User receives feedback based on their response (both traffic and debris are hazards in the image).



Examples of more detailed/technical knowledge introduced in the study.



Another “check your knowledge” question.



This is an example of a more challenging question for a similar instructional context.

Adaptation can also include the choice of images with more/less challenging perceptual content. The dog (potential hazard) is easier to perceive in this image than the following one.

Subjects can also be asked to identify specific areas on an image corresponding to an instructional concept (in this case, identifying hazards).


APACTS -
Scene Siz
Cancel
Identify all visible hazards in the image. Press the arrow on the right to move on.

Appendix E

Recruitment Flyer

Have you ever wondered
what it takes to become an
Emergency Medical Technician?

Participate in our study to find out!

We are running a study to test a new technology for training EMTs. In our study, you will learn some of the knowledge required for EMTs, then apply what you've learned in an interactive learning environment.

You will be compensated $15 for participating.

Please contact Kimberly Stowers at kstowers@cba.ua.edu for more information.

APPENDIX C: Exploration of Behavior Markers to Support Adaptive Learning

Exploration of Behavioral Markers to Support Adaptive Learning

Adam Wearne, Robert E. Wray

Soar Technology, Inc. 3600 Green Court Suite 600, Ann Arbor, MI 48105 USA
{adam.wearne, wray} @soartech.com

Abstract. In designing and developing adaptive learning systems, it is desirable to incorporate as much information about the learner as possible to better tailor in instructional experience. Behavioral markers exhibited by the learner offer a source of information with the potential to shape instructional content. In the case of computer-based training environments, this source of information may include behaviors ranging from mouse cursor movement, key stroke dynamics, or eye tracking. We present methods for analyzing the mouse behavior of a learner using kinematic data in situations where knowledge of areas of interest on the screen are not known by the system a priori, as well as in multiple-choice scenarios to analyze the amount of attention spent by the user on various response items. The outcome of this work is to help inform and to influence future studies in adaptive learning which may seek to incorporate such sources of learner information.

Keywords: mouse tracking, adaptive training

1 Introduction

Incorporation of behavioral data has become an increasingly active area of study across a wide variety of domains. Search engines and online retailers have a great deal of interest in understanding how users interact with their web pages [1, 2]; keyboard dynamics and typing patterns have been used as an additional mechanism for account security [3]; and research results in eye-tracking have found applications in brick-and-mortar retailers in an attempt to improve overall customer experience [4]. Inclusion of behavioral data into the domain of student learning may have the potential to improve tailoring of pedagogical content [5].

The ultimate goal of this work is to understand and to respond effectively to end-user behavior in such a way that training content is more optimally tailored to that student's needs. Increasingly, training scenarios tailor content based on student responses to situations and their performance on trials and tasks. A potential improvement to this paradigm is to include alternate sources of information to aid in the tailoring process. For

instance, understanding where a student's attention is focused may help us to understand possible sources of student confusion, and allow us to better tailor instructional content to remediate specific learning objectives or increase the training difficulty. Some potential vectors for behavioral markers include eye-movement, keystroke dynamics, and mouse cursor movement. This paper focuses on mouse-cursor dynamics and their influence on student learning.

We provide a brief review of recent work in this field with a particular focus on mouse-cursor tracking and the attempts that have been made to do trajectory and target prediction, as well as post hoc analysis of cursor activity. Much of the work in the predictive domain focuses on applying Kalman filtering techniques [6], neural networks [7], and kinematic models to raw mouse trajectory data [8]. Post hoc analyses tend to emphasize the aggregate statistical properties of user behavior. We then present a novel methodology for analyzing and classifying a user's mouse activity according to a set of behavioral heuristics suited for computer-based training. Results display data gathered on raw kinematic information for target recognition and sequence tracking which are aimed at identifying user velocity and acceleration information. Additionally, we examine user responses on a simple multiple-choice quiz to investigate dwell times and elapsed time between target identification and initial click.

Building upon this work, future goals are then to apply this framework to a student study in which participants are presented with training material on a specific topic and remediated based on both their raw response data as well as mouse tracking information. Further extensions may include eye tracking via either webcam or specialized devices, as well as real-time analysis of keystroke dynamics.

2 Mouse Tracking as a Behavioral Marker

Predictive mouse tracking is an active area of research with myriad applications. Gaining a deeper understanding of how users interact with devices, and how they interact with elements on computer screens is crucial to developing systems with an improved user experience.

Biswas and Langdon [7] developed a neural-network based method for determining user intent in an attempt to improve the user experience for individuals with severe motor impairment. The network separates a single mouse movement into two main phases. First, a "homing" phase denotes the period during which the user is deciding upon an intended target. After the homing phase is identified, the system then predicts regarding the user's intended target. The network takes as input the velocity and acceleration of the mouse cursor, as well as the bearing angle with respect to all target locations on the screen. The network is trained such that it can recognize the "homing" phase of a mouse cursor

trajectory. When the network predicts that the mouse cursor is within the homing phase, the system identifies the intended target by examining which screen elements are closest to the position of the mouse cursor, and which are also in line with the bearing angle of the cursor trajectory. The identified screen element is then enlarged and highlighted in an attempt to assist the user in more easily selecting the desired target. This approach is a successful effort in the reduction of pointing times, and has found particular use in assisting those users with motor-impairments interact with their devices more easily. However, this approach requires knowledge of screen elements which the user may interact with for the purposes of calculating the bearing angles and other kinematic information required to make predictions using their neural network architecture.

Pasqual and Wobbrock [9] explored a template matching approach based on kinematic data to estimate the endpoint of a mouse stroke gesture. In this work, a repository of historic trajectories was created and used as a basis for comparison on new stroke gestures previously unobserved by the system. Using the elements of this repository as templates, one then computes the closest matching template motion based on a scoring heuristic to match the new data sample to the most similar template. Using the total distance travelled by the selected template, the end point of the incoming gesture is estimated to be the same pixel distance away from the starting point in the direction of motion of the gesture. This work is novel in the sense that it provides a means for endpoint prediction in a "target-agnostic" sense, in contrast to methods which require knowledge of screen elements and their locations relative to the cursor.

3 Kinematic Analysis

One of the common ways in which analytics is done on cursor activity is via analysis of cursor kinematics. To investigate this approach, a sample task was created in which users are prompted to click on certain screen elements – occasionally in a specific order. From this sample task, we can then track the user's cursor across the screen as it moves between various elements of interest and attempt to understand, at some level, the user's intended target based on raw kinematic information.

To extract useful information from the raw data, we base our method on Fitts' Law [10]. That is, as the user's cursor approaches its intended target, the velocity of movement will begin to decrease. With this model in mind, efforts focus on estimating the local maxima of cursor velocity. We can capture this information then by processing the raw input data of screen coordinates to understand the time at which the user has 'locked on' to a given target. The raw input data consists of time stamps and screen coordinates. Given this, it is trivial to calculate velocity components, as well as the overall magnitude of velocity.

$$v^2 = \left(\frac{x_{t+1} - x_t}{\Delta t}\right)^2 + \left(\frac{y_{t+1} - y_t}{\Delta t}\right)^2$$

However, due to the granularity of the data, this estimate is very noisy. To address this concern, we employed a two-stage smoothing method. The first stage consists of a simple moving average of the velocity data. This is done as a first pass at removing some of the short-term fluctuations of the velocity data associated with micro-adjustments of the cursor position. To further smooth the data, we then apply a low-pass filter on the moving average data to improve the signal-to-noise ratio of the velocity data. The filter removes high-frequency artifacts such as mouse "jiggle" that can occur during mouse movements.

Figure 1 illustrates the behavior of a user on a sample task in which they are asked to click on various screen locations. Note while the user is directed to certain locations, the underlying system has no knowledge of the actual spatial layout of screen elements of interest. The blue/green curves indicate the position of the cursor over time, and the black dashed vertical lines indicate the times at which the user moved on to the next question. The periodicity observed in this time series is a result of the fact that the user must click in a specific region of the screen to advance to the next task. The corresponding cursor velocities are displayed in the lower portion of Figure 1. By identifying the peaks of the velocities, we are essentially identifying the moment at which the user's cursor begins to decelerate and hone in on the target location.

As a concrete example, consider a particular task in which the user is asked to interact with screen elements in a particular sequence, as illustrated in Figure 2. What is noteworthy about this specific task is that the most intuitive cursor path to accomplish this would have the user pass through elements of interest, which are meant to be interacted with later in the sequence. More elementary methods which only detect intent based on bounding boxes surrounding elements of interest would then fail by 1) over-counting the number of intended actions of the user and 2) misrepresenting the temporal order of user intentions. The raw cursor trajectory of a user on this task is displayed in Figure 2. Temporal ordering of cursor location is represented by a color gradient ranging from yellow (start of the task) to red (end of the task).

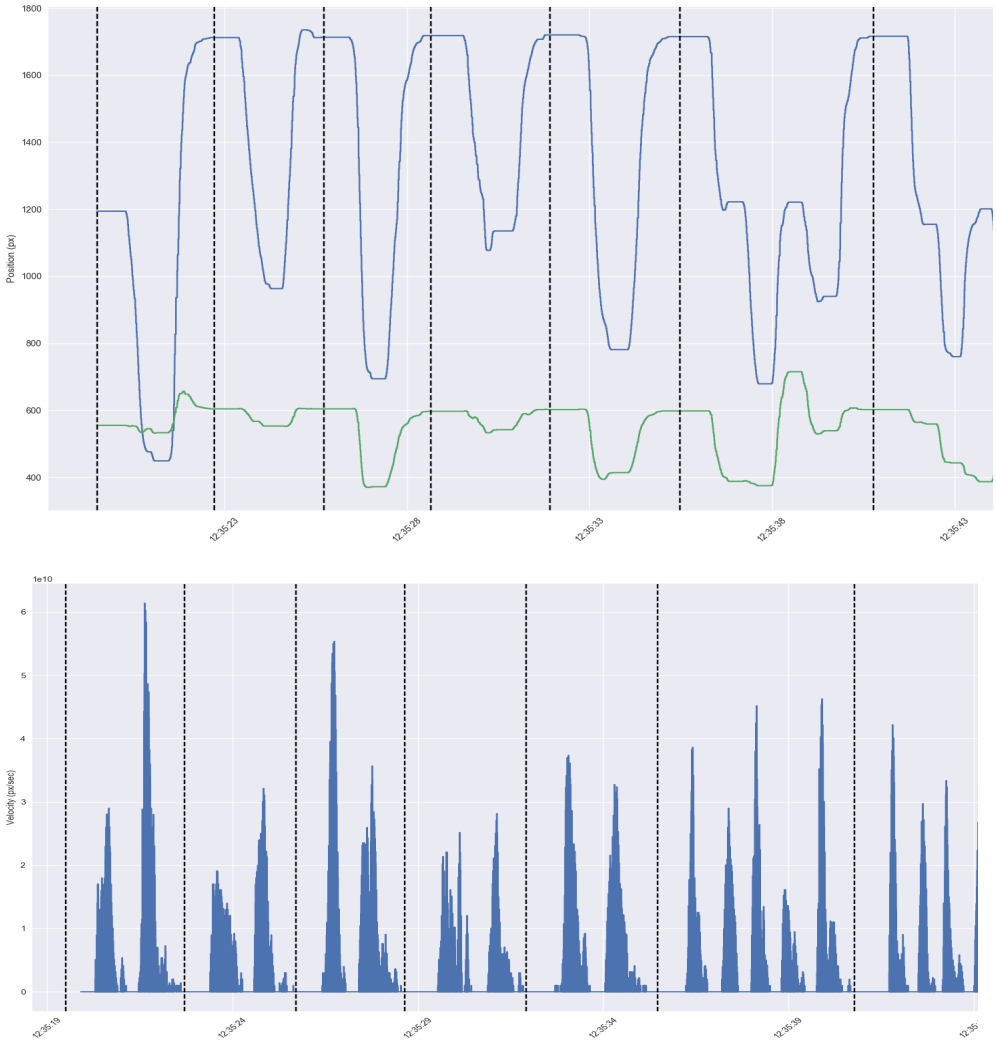


Figure 1: Screen coordinates for each activity in sample task (top). Magnitude of cursor velocity for each frame (bottom). Vertical dashed lines denote individual frames of sample task.

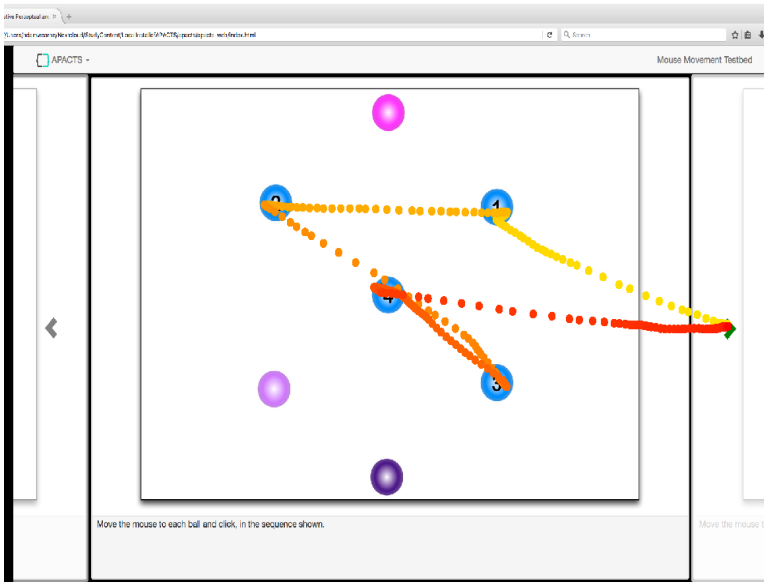


Figure 2: Raw cursor screen information for a given sample task.

Taking in the raw kinematic data from this particular task, we must first smooth the data perform attempting to identify the velocity peaks. This is a necessary step to avoid finding many local maxima in the raw data, which would result in many timestamps being erroneously identified as a peak. This is done by application of a Savitsky-Golay filter [11], a low-pass filtering method to improve the signal to noise ratio. After smoothing the data, we then can then detect the peaks of the system using standard numerical software packages. The results of such a procedure as applied to the raw cursor movement data represented in Figure 2 are displayed in Figure 3. Using estimations of the local gradient to find local optima, a total of six peaks were identified. Four of these peaks correspond to the locations of interest that are relevant to the task. The additional points are artifacts of the user selecting the green arrow on the right-hand side to advance to the next frame.

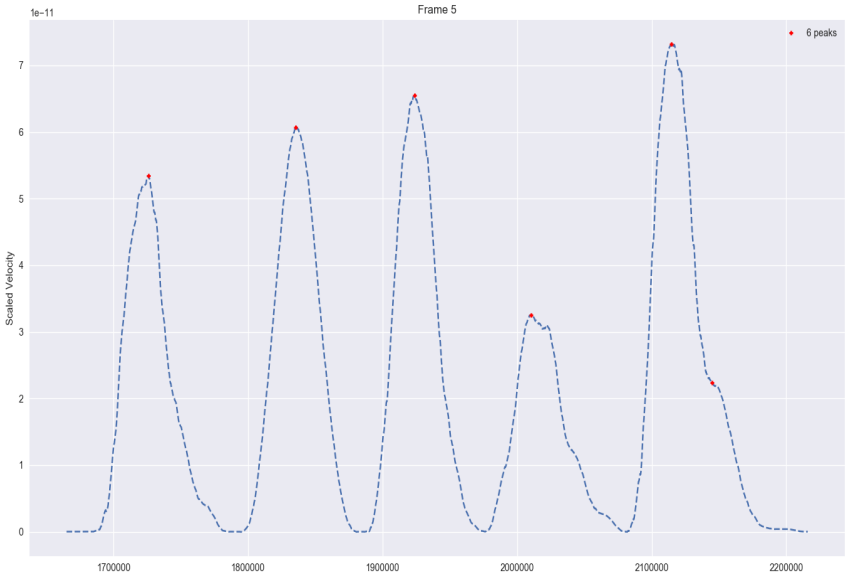


Figure 3: Filtered velocity curves with identified peaks.

Using the timestamps of these identified peaks, we can then filter the original raw trajectory data and visualize the locations and times at which the user's intent was recognized under this system, as illustrated in Figure 4. There are several features of this result worth noting. First is the fact that the system is able to identify all the intended elements of interest on this trial task without having any knowledge of their location. Such an ability may facilitate the design process of research projects and studies by helping to alleviate (at least at some level) the encoding of bounding boxes for elements of interest on the screen. Further, we can see the system is able to identify the correct temporal ordering of intended actions of the user as demonstrated by the color gradient of the scatter plot markers. Combining this with additional domain specific information may then allow for prediction of user intent.

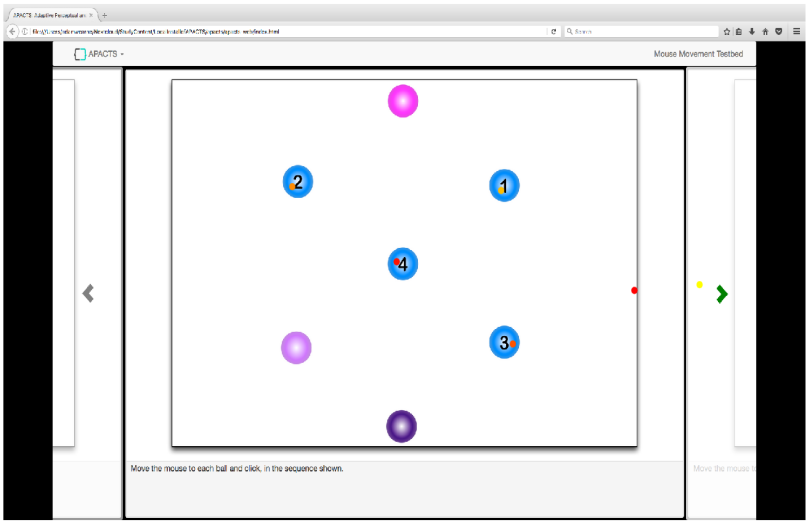


Figure 4: Identified locations of user attention based on filtered velocity data.

4 Multiple-Choice Domain

In addition to examining user behavior on more open-ended tasks like those mentioned in the previous section, we also investigated mouse cursor information in domains in which the user's expected range of motion is much more limited: multiple-choice quizzes. In contrast with the above, the objective is slightly different. Instead of attempting to make potentially predictive statements regarding the user's intent, we perform post-hoc analysis of user behavior across a series of multiple-choice questions. The goal is to identify what choice(s) the user considered, and to estimate user confidence when responding to a given question. Such information could then be used in concert with a dynamically-tailored instructional component to potentially improve the quality of training a user receives by remediating the learning process specific to the user's individual needs.

For the purposes of categorizing user behavior in the multiple-choice domain, analysis focused on behavioral markers related to the user's interaction with screen elements: 1)total time spent on a given question, 2) amount of time the user's cursor spent hovering over each multiple-choice option, and 3) time elapsed between the cursor hovering over an option and clicking. Using this behavioral information in addition to score data associated with each multiple-choice response, we have developed a heuristic framework that classifies the user's behavior on a given question as falling into one of several categories. A brief description of each of these categories is listed in Table 1.

Table 1. Behavioral categories and descriptions.

Category		Description
1	**Checkout Type I**	Learner makes a rapid, high confidence choice before there is sufficient time to interpret the content of the questions.
2	**Checkout Type II**	Learner chooses a "checkout" option from the question options.
3	**Mapping Error**	Student makes a selection that indicates a misunderstanding of the problem (categorization error)
4	**Error of Omission**	For combo questions, learner chooses a response that leaves out one of the correct choices.
5	**Error of Commission**	The user chooses an incorrect response from a multiple-choice question.
6	**Close Call**	User weighs several options including the correct one, but ends up making a wrong choice.
7	**Lucky Guess**	The user chooses a correct answer but appears to have little confidence in that answer.

After the user has been presented with a given multiple-choice question and responded, their actions on that question are analyzed in the following way. For each available choice on a given question, let $m_i \in \{m_1, m_2, \dots, m_K\}$ denote the set of responses where K denotes the total number of responses. The amount of time spent on this frame denoted as T. We can then construct a list of the user's behavior on a given question by creating a series of tuples corresponding to the actions taken by the user with respect to each response. This series of tuples takes the form:

$$\left\{\begin{pmatrix} a_1 \\ t_1 \\ c_1 \end{pmatrix}, \dots, \begin{pmatrix} a_N \\ t_N \\ c_N \end{pmatrix}\right\}$$

Where N is the total number of unique visits to any of the response choices, a_i is the response visited on the i-th step, $t_i$is the amount of time the user's cursor was hovering over a given choice, and $c_i$is the time elapsed between hovering over the response and clicking on it. Note that when specified in this way, the system is subject to the constraint that $\sum_{i=1}^{N} t_i \leq T$.

Given this series of observations, we can then construct a measure of consideration for each choice based on the amount of time spent hovering over a given response, and how many times the user may have selected that response. We calculate the partial attention score for each response by taking the total non-trivial amount of time spent on each response. That is, for choice m_i,

$$\text{Partial Attention}(m_i) = \sum_i \max(t_i - \alpha, 0)\ \delta(m_i, a_i)$$

Here, α is a parameter that controls for what dwell times are considered non-trivial. Here, we also introduce a kronecker delta to ensure that we are only taking the sum over the relevant tuples within our series of observations.

To estimate the total amount of attention spent by a user on each response choice, we also incorporate the information from the click deltas. In a similar manner, we can augment the partial attention score by an additional term proportional to the total amount of elapsed time between hovering and clicking a given response. This has the effect giving heavier weight to responses that were selected by the user. Such a weighting scheme also makes intuitive sense because clicking on a particular response provides a much stronger signal of user attention that hovering does on its own. The total consideration score (TCS) afforded to each response choice is then given by

$$\mathrm{TCS}(m_i) = \frac{1}{T}\left(\text{ Partial Attention}(m_i) + \sum_i \max(c_i - \gamma, 0)\delta(m_i, a_i)\right)$$

Where the model parameter γ controls for non-trivial click delta times. The final score is then normalized according with respect to the total amount of time spent on that question.

Given the vector of TCSs that have been awarded to each response for the question, the vector is then reordered according to the magnitude of each element's TCS. To determine if the user paid a similar amount of attention to multiple responses, we then take the percent difference in TCS between neighboring elements of the reordered TCS vector. If the magnitude of this percent difference is below some threshold, then we regard the two choices as having been considered a near equal amount by the user.

Using this information, we can then discriminate between the potential user behavior categories as defined in Table 1. In the simplest case of Checkout Type I, the amount of time the user spent on a given question is compared with the median amount of time spent as measured by some sample users. If the user spends much less time on a given question compared to this median value, then the system indicates that the user has not spent a sufficient amount of time on the question to consider or interpret the material presented to them. For more nuanced categories like the Close Call, we consider if the percent differences as calculated by the associated TCS vector indicate if the correct response was considered even if the response the user ultimately selected was incorrect. This is in contrast to behaviors typified by Mapping Errors in which the user may not have considered the correct response, and instead weighed several incorrect options. In the case of Close Calls, the system has some confidence in the user's understanding of the material given that the correct choice has a comparable TCS relative to the selected one. In the case of Mapping Errors, the user has not sufficiently demonstrated this understanding and may require additional instruction. One can see then, by capturing these behavioral markers that are often overlook, one may be able to more finely tune instructional

remediation to better fit the specific needs of the learner, whether it be a quick reminder about instructional content, or a more substantial review of the material under study.

To estimate the free parameters of our model, we examined the behavior of several participants on a simple multiple-choice quiz. Using behavioral information gathered on a mock quiz, we can then create an initial anchoring for the values of the model parameters. The results of this small-scale study are displayed in Figure 5.

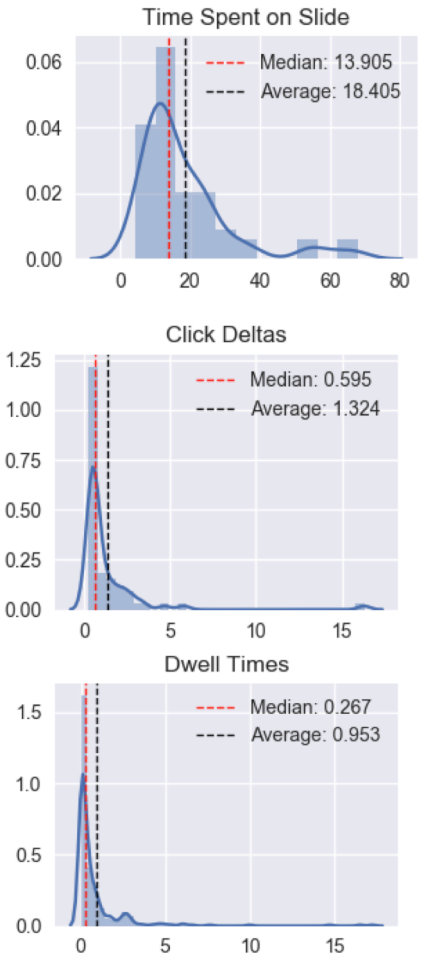


Figure 5: Distributions of model parameters gathered from trial study.

5 Conclusion

In this paper we outlined procedures for analyzing mouse cursor behavior from the perspective of both raw kinematic data, as well as from more stylized data formats more apropos to what one might expect to see on a multiple-choice test. The kinematic analysis described here provides a simple yet effective method of uncovering the user's attention through a process of filtering velocity information and peak detection. Such a method may assist researchers in understanding a user's intent even in cases where the computer-based training system does not have access to information regarding the location of individual screen elements. The framework laid out for behavioral recognition of user actions in multiple-choice scenarios is also noteworthy in that it provides an inclusive framework for interpreting and discriminating between several categories of user actions with minimal need for historic data.

Our future plans include incorporating this framework into a study regarding the impact that behavioral markers may have when they are including as part of the instructional

tailoring process. Further work includes large-scale analysis of user behavior to classify user behavior from a data-driven approach. While the heuristic method described in this article presents a first attempt at the problem of understanding behavioral cues of users, we would like the ability to employ machine learning classification techniques to better understand how users are responding. Larger scale studies will provide data to support development of more data driven solutions (such as classifiers derived from machine learning). These new solutions can then be used in subsequent studies to refine mouse-based behavioral markers for adaptive learning systems.

Acknowledgments.

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs, through the Joint Program Committee-1/Medical Simulation and Information Science Research Program under Award No. W81XWH-16-1-0460. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense. The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office.

References

1. C. Tenopir, P. Wang, Y. Zhang, B. Simmons and R. Pollard, "Academic users' interactions with ScienceDirect in search tasks: Affective and cognitive behaviors," Information & Process Management, vol. 44, no. 1, pp. 105-121, 2008.

2. G. Moody and D. Galletta, "Lost in Cyberspace: The Impact of Information Scent and Time Constraints on Stress, Performance, and Attitudes Online," Journal of Management Information Systems, vol. 32, no. 1, pp. 192-224, 2015.

3. R. Ponkshe and V. Chole, "Keystroke and Mouse Dynamics:A Review on Behavioral Biometrics," International Journal of Computer Science and Mobile Computing, vol. 4, no. 2, pp. 314-345, 2015.

4. E. Bradlow, M. Gangwar, P. Kopalle and S. Voleti, "The Role Big Data and Predictive Analytics in Retailing," Journal of Retailing, vol. 93, no. 1, pp. 79-95, 2017.

5. P. Ottavi, M. Pasinetti, R. Popolo, G. Salvatore, P. Lysaker and G. Dimaggio, "Chapter 17 - Metacognition-Oriented Social Skills Training," in Social Cognition and Metacognition in Schizophrenia, San Diego, Academic Press, 2014, pp. 285-300.

6. G. A. Aydemir, P. M. Langdon and S. Godsill, "User Target Intention Recognition from Cursor Position Using Kalman Filter," Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion, 2013.

7. P. Biswas and P. Langdon, "Multi-modal Target Prediction," Universal Access in Human-Computer Interaction. Design and Development Methods for Universal Access, 2014.

8. B. D. Ziebart, A. K. Dey and J. A. Bagnell, "Probabilistic Pointing Target Prediction via Inverse Optimal Control," Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, 2012.

9. P. T. Pasqual and J. O. Wobbrock, "Mouse Pointing Endpoint Prediction Using Kinematic Template Matching," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014.

10. P. Fitts, "The Information Capactity of the Human Motor System in Controlling the Amplitude of Movement," Journal of Experimental Psychology, 1954.

11. A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.," Analytical Chemistry, 1964.

APPENDIX D: Assessing the Role of Behavioral Markers in Adaptive Learning for Emergency Medical Services

Assessing the Role of Behavioral Markers in Adaptive Learning for Emergency Medical Services

Kimberly Stowers[1], Lisa Brady[1], Youjeong Huh[1], Robert E. Wray[2]

[1]The University of Alabama, PO Box 870225, Tuscaloosa, AL 35487, USA
kstowers@cba.ua.edu, {llbrady, yhuh1}@crimson.ua.edu
[2]Soar Technology, Inc., 3600 Green Court Suite 600, Ann Arbor, MI 48105, USA
wray@soartech.com

Abstract. Tools for adaptive learning are on the rise, resulting in the creation and implementation of increasingly intelligent tutoring systems. These systems can be applied in a variety of contexts, including civilian, military, and emergency operations. Such systems may also include the capability to adapt to learner needs based on performance or behavioral input. However, the use of such adaptation may vary in its success depending on the domain it is applied to. This paper examines the potential utility of adaptive tutoring for educating Emergency Medical Service (EMS) workers. We examine two complementary approaches that can be used to drive adaptation: performance-based and behavior-based adaptive learning models in intelligent tutoring. We then discuss implications of implementing such learning models for intelligent tutoring in EMS. Next, we outline ongoing research as a use case for the validation of different adaptive learning models. Finally, we discuss expected impacts of this line of research, including the expansion of adaptive tutoring to other domains related to EMS.

Keywords: Intelligent tutoring · Training Design · Adaptive Training

1 Introduction

The world of work is changing. A 2016 Pew Research Center [1] survey found that 54% of workers believe it will be essential for them to develop new job skills throughout their work life to keep up with changes in the workplace. Such beliefs are well-founded, as many are forecasting that the rapid proliferation and advancement of technology will cause many changes in the employment landscape across a broad range of occupations and industries, (e.g., [2]). Certain fields, such as emergency medicine, are increasingly faced with a number of changes related to advanced technology and equipment, modified regulations, and updated safety procedures. Because the field is always changing, emergency personnel must constantly be re-educated to not only learn the changes but also implement them into their routines.

Although having students receive individualized instructor assistance during any learning process is ideal, the reality is that these resources rarely exist. One solution, however, is the development and utilization of intelligent tutors (i.e., computer-based artificial agents [3]). Intelligent tutors may not only approach the effectiveness of

human teachers, but also be able to customize and personalize instruction to achieve efficient and effective learning outcomes for all students. The purpose of this paper is to examine the utility of intelligent tutoring for Emergency Medical Services (EMS). We begin by reviewing the state of the science of adaptive learning—including performance-based and behavioral-based adaptive learning models—as well as their propensity for use in intelligent tutoring for EMS. We then present ongoing research as a case for the validation of related adaptive learning models. Finally, we discuss the impacts that can be expected as a result of designing and implementing adaptive tutors for EMS and across a variety of occupations.

2 Adaptive Learning for Emergency Medical Services

Adaptive learning, in which a learning environment adapts to the abilities, needs, and preferences of individual learners, has been identified as a "Grand Challenge" for 21st century research and engineering [4]. The benefits of adaptive learning environments include more efficient learning [3], improved attention and motivation [5], the development of less rigid and more flexible decision making (i.e., adaptive expertise, [6]), and improved transfer of learning to settings in which learned knowledge is used and applied [7-9]. Adaptive learning has application for more pervasive and less costly training, as opposed to training that has traditionally been delivered by human instructors in classes with modest student-to-teacher ratios. Human instruction is very valuable and—in some cases—can be adapted to individual learners in a meaningful way. However, technological adaptive methods offer an opportunity to deliver effective training tailored to a greater degree and available to a greater number of students. This, in turn, can enable a single human instructor to train many more students with a greater degree of individualization.

In the context of EMS training, adaptive learning may lead to greater training and performance outcomes, especially if the tutor in question utilizes scenario-based learning methods (e.g., [10]) wherein the content to be learned is presented in the context of the real-world environment [11,12]. Indeed, research has shown that EMS workers' performance in a simulated field environment matches their success in completing standardized EMS tests [13]. Additionally, results from a prior study assessing the utility of a serious game for training EMS nurses suggested that adaptation of training to varying levels of expertise might be an improvement over a static (i.e., non-adaptive) virtual training environment [14]. What remains unknown, however, is how best to apply adaptive methods to an EMS training environment.

2.1 Approaches to Adaptive Learning

Adaptation to a learner usually requires a model of the learner that is frequently updated as a learner progresses through a curriculum [15]. The targeting of adaptive techniques, such as scaffolding [16] and competency matching [17,18] depends on the accuracy (and, to some degree, precision) of the learner model. When the model better reflects the learner's actual knowledge, skills, and attitudes at any point during the learning, the targeting of the adaptive method to the learner generally improves [18].

Creating a complete and accurate learner model is difficult, however, and several approaches exist. Traditionally, adaptive learner models were created from formal and informal assessments within the learning environment [15, 19-21]. Over time, adaptive learning methods have evolved to include other triggers for adaptation, including "markers" of human behavior and biology related to cognitive states in the learner. Next, we discuss these two approaches for creating adaptive learning systems. Specifically, we look at the utility of performance markers and behavioral markers for adapting EMS training to individual learner needs.

Performance Markers. Performance-based adaptation focuses on altering learning content according to the learner's performance over time. This may include giving immediate feedback to the learner (e.g., [19]), redirecting the learner to content that should be reviewed to improve or maximize performance (e.g., [22]), or choosing what content is most appropriate to present next as learning progresses [23]. Additionally, the learner may be able to repeat or review content as desired. Regardless of the particular approach, to maintain ongoing adaptiveness, the tutor should track learner performance over time. Thus, several performance markers have been identified to aid in this process.

Performance markers can be thought of broadly as markers that represent the knowledge, skills, abilities, and other relevant characteristics (KSAOs) of the learner [24]. By linking KSAOs to learning or training objectives, it becomes possible to identify domain-specific performance markers. For example, what KSAOs should the learner demonstrate throughout training to indicate s/he is meeting a particular learning objective?

Linking performance markers to learning objectives provides ample opportunity to create an adaptive tutor that tracks performance effectively in EMS training. In the domain of EMS, learning objectives vary according to the specific position of interest. For example, the National Standard Curriculum for Emergency Medical Technician (EMT) training divides learning objectives into three categories--cognitive, affective, and psychomotor--and three levels--knowledge, application, and problem-solving [25]. Identifying performance markers that represent some or all of these categories and levels may prove useful to the adaptation of the training. However, for computer-based training of real-world skills (such as patient triage), it is often difficult to develop accurate performance-based markers that represent true operational knowledge. Additionally, the method of assessment can be confounding as well. As a simple example, when a learner is asked to answer a multiple-choice question, a correct answer could be indicative of actual knowledge and ability, or it could be due to other factors (lucky guess, lack of time pressure, access to other resources, etc.) Rather than use performance-based markers alone, these limitations suggest the need for complementary functions that can mitigate some of these limitations.

Behavioral Markers. In addition to estimating learner capability from formal and informal assessment within the environment, researchers have explored many behavioral, physiological, and even neurological indicators or "markers" that can provide additional context for improving the dynamic assessment of the learner. The purpose

of such markers is to identify learners' cognitive states and select learning strategies based on that information [26].

Learner states of interest may include arousal, attention, and other cognitive states or processes that impact learning behaviors [27-28]. In particular, understanding the dynamic patterns of learner arousal allows the identification of dynamic adaptation targeted to the identified arousal states [29]. To that end, markers that have been used to track learner arousal/attention in learning environments include behavioral sensors (e.g., posture, expressions, mouse movements), physiological sensors (e.g., galvanic skin response [GSR]), and neurological sensors (e.g., electroencephalography [EEG]; and see [30]).

Tracking learner states may be particularly useful for guiding learner participation in EMS training. Indeed, since a training program is only as successful as its ability to engage learners [31], implementing a system that can track cognitive states and respond to those states in a way that increases engagement and learning is well-advised. Prior research examining attitudes of EMS students toward computer-based training showed that preferences for such training were moderate at best, suggesting that lack of engagement might dampen the students' attitudes [32]. Additionally, because some aspects of successful EMS performance includes the ability to handle pressure and stress, a learning environment that could appropriately raise learner arousal (while not overwhelming the learner) would be beneficial for deeper learning. As such, adaptation that uses behavioral markers to facilitate engagement and personalization beyond a traditional learning environment may be warranted.

2.2 Selecting Markers for EMS Training

In the domain of EMS training, a combination of both performance and behavioral markers may be most promising for creating an adaptive learning system that is both accurate and practical. As stated above, the selection of performance markers should be linked to the objectives in question. A straightforward performance marker may simply be the "correctness" of the learner's response. However, the marker should be further contextualized according to the materials being learned. For example, contextualizing a marker to fit the National Standard Curriculum for EMT [25] would involve marking the response according to the category (i.e., cognitive, affective, psychomotor) and level (i.e., knowledge, application, problem-solving) of questions answered. This would then allow the learning model to take into account whether the learner is doing better in certain categories than others. Similar approaches can be taken for other EMS curricula.

EMS training can be further personalized through the implementation of meaningful behavioral markers. However, most behavioral markers today require sensors that are not commonly available on the hardware available for typical computer-based learning: a laptop or a tablet. This can make the implementation of behavioral markers quite costly. In addition to the expense of developing tutors that implement various behavioral markers, supplying the hardware necessary--especially in learning environments that may not otherwise have them--may be infeasible or unrealistic.

Thus, it is important to consider ways in which behavioral markers can be implemented at a low cost and with low intrusion to available learning environments. One such marker that can be readily implemented in a computer-based environment for a

fairly low cost with little disruption is mouse-tracking. The utility of mouse-tracking lies in its ability to infer learner focus during a decision-making process. For example, in the case of a learner viewing a multiple-choice question, mouse-tracking can provide real-time evolution of the decision making process the learner is engaging in, including which responses s/he is considering most heavily [33]. For EMS, there may be further use in tracking mouse movement as learners view specific images in which they are asked to identify emergency-related factors.

To that end, combining mouse-tracking with contextualized performance tracking can provide insight to the learner's evolution of knowledge and skills during a learning session, as well as the process underlying that evolution. Furthermore, combining these approaches can both contextualize their use as well as provide further validation for their utility in adaptive learning.

3 Ongoing and Future Work

To explore recommendations discussed throughout this paper, we have developed an adaptive learning system for EMS which has the capability to exploit both performance markers and behavioral markers. The basic learning environment has been used to present computer-based learning and scenario-based practice in a web-delivered system [34, 35]. We have extended the environment to perform adaptation based on markers derived from mouse-tracking [36] to complement the adaptive performance based markers built into the tool [34].

A previously developed analysis provided insight to the role and impact of markers for a well-designed curriculum for EMS [37]. Curriculum units have since been developed for "Scene Size-Up," a required curriculum component used in Emergency Medical Technician (EMT) training [25]. Scene size-up regards the initial assessment of the scene by the EMT on arrival and spans all of the learning goals outlined for the curriculum above. Using these units, a study has been designed to identify the impacts of performance-based adaptation on learning and whether (and to what extent) there is benefit for using behavioral markers in addition to performance markers in this adaptive EMT training.

The study, currently underway, compares the results of learning in a scene size-up presented in three different ways: two adaptive units, and a unit presented in a non-adaptive or fixed sequence. The three variations in the adaptive tutor being tested can be described as such:

- Condition 1: Non-adaptive. The instructional unit is presented in a fixed sequence. Learners receive generalized feedback to every response (both correct and incorrect responses).
- Condition 2: Adaptive based on performance (only). The instructional unit is presented in a fixed sequence that includes immediate, adaptive feedback in response to learner questions. Learners receive feedback specific to their response (correct, incorrect) and targeted to the specific choices the learner made. Highly specific feedback and remediation is provided for "close" responses; more general, conceptual feedback is given for responses that are (conceptually) far away from the target response. Such feedback design is consistent with guidance for the design of feedback delivery in learning sys-

tems [38]. Learner knowledge and skill is estimated and updated as learning progresses.

- Condition 3: Adaptive based on performance and markers. The instructional unit is presented in a variable sequence. The sequence and feedback are dynamically constructed/chosen based on a combination of direct learner observation (as above) and behavioral markers—specifically indicators captured via mouse-tracking. Only two types of sequence variation are introduced in this study: repeating a prior question and skipping a question.

The primary performance marker tracked in this study is response selection (i.e., correctness of selected response) which is mapped to the learner model. As stated above, mouse-tracking is the behavioral input. We have developed the following markers for this study based on the mouse tracking input [36]:

- Confidence: An assessment of the learner's confidence in a selection or choice. The marker is derived from a combination of movement patterns and dwell times on screen items.
- Secondary choice: An evaluation of learner choices that attempts to identify if the learner considered another choice (or subset of all choices) more heavily than the other choices.
- Likely target: For item selection tasks, this marker predicts a learner's likely mouse target in advance of reaching that target, reducing the chance that mouse overs that occur during movement to a target are treated as targets.

These markers are used to make adaptive selections of content, feedback, and remediation in condition 3. For example, a learner that selects the correct choice with low estimated confidence may be given some feedback and remediation following that choice while a high confidence learner is allowed to proceed immediately to the next item. Similarly, a learner that selects a "far target" choice with high confidence is immediately asked to try again without any feedback under the assumption that this choice may be due to lack of engagement than lack of knowledge.

The primary outcome that will be assessed in this study is knowledge gain, quantified as difference scores between pre- and post-tests given to participants. The study will be conducted with university participant population(s), which will allow for a sufficient sample to assess the utility of each tutoring approach. This will provide a strong foundation for future studies to pursue further validation with incoming EMT and other EMS worker populations to indicate their success with its use.

4 Expected Impacts

Although the primary focus of the present chapter is examining the utility of adaptive tutoring for educating EMS workers, furthering research in this area can be expected to have broader impacts than those discussed thus far. Specifically, these impacts include: 1) improving the state of the science on adaptive tutoring; 2) providing optimal and affordable EMS training; and 3) extending the use of adaptive training from EMS to other occupations.

4.1 Improving the State of the Science on Adaptive Tutoring

As discussed prior, adaptive learning has been identified as a "Grand Challenge" for 21st century research and engineering [4]. Improved adaptive learning has applications for more pervasive and less costly training in a wide variety of domains, from classroom education to on-the-job training. For example, research has been conducted in the context of training soldiers in the U.S. Army using an adaptive tutor to make one-to-one tutoring possible (e.g., [26]). However, challenges still remain regarding the cost of authoring effective adaptive tutors, as well as discovering adaptive tutoring technologies that can more accurately perceive a learner's state and progress [26]. To that end, testing the utility of various behavioral markers in combination with performance markers can aid in designing more precise learner models for adaptive tutoring systems.

4.2 Providing Optimal and Affordable EMS Training

One of the challenges EMS practitioners are currently facing is a number of changes on the horizon in emergency medicine, such as modified regulations and procedures. The constantly changing nature of EMS protocols requires all staff in various levels to be continuously re-educated. One of the most effective solutions to this need is to design adaptive tutors that move beyond simply identifying individual errors and advance toward understanding the process and needs involved in knowledge acquisition. Adaptive tutors can provide fine-tuned, individualized training, making it easier to educate people with varying levels of expertise. In addition, from a practical standpoint, utilizing low-cost, low-intrusion behavioral markers will allow EMS personnel to receive a unique, adaptive training that is less costly and more convenient than neuro-cognitive markers.

4.3 Extending the Use of Adaptive Tutoring to Other Occupations

Although EMS personnel receive training that is specific to their role, there are a number of occupations in which professionals (e.g., police officers, fitness trainers, and lifeguards) are required to learn similar skills—such as those related to CPR and First Aid. Designing adaptive tutors for EMS can thus aid in designing specialized learner models for occupations across a variety of industries. Although additional research is needed to validate the use of markers and learning models that are relevant to on-the-job training in other domains, the implementation of adaptive tutoring within training contexts is likely to increase dramatically in the near future. Due to imminent changes within the workforce, employees within a variety of industries will need to acquire additional knowledge, skills, and abilities. Thus, designing specific adaptive tutoring platforms that respond to the training needs within each industry could provide a solution to this widespread need.

5 Conclusion

The workplace continues to evolve rapidly as a result of technological, societal, and political changes, making our discussion on the development of intelligent tutors timely and essential. Although current and future employees acknowledge the need to constantly learn and apply new skills to perform well in their role, a lack of time and resources makes this process difficult, if not impossible. With the increasing use of mainstream adaptive learning tools, however, intelligent tutoring may provide a solution. In this paper, we evaluated the science of performance-based and behavioral-based markers used for adaptive learning, discussed the utility of creating an adaptive intelligent tutor that incorporates meaningful behavioral and performance markers of learning to train EMS personnel, and provided an example of ongoing research within the specific domain of EMT. Our examination of strategies involved in creating adaptive tutors provides an example of these tutors can be used within a specialized and high-risk occupation. However, as the evolving world of work begins to generate related changes within training environments across a variety of industries, the use of adaptive tutors may be not only helpful, but necessary.

Acknowledgments. This work was supported, in part, by the Office of the Assistant Secretary of Defense for Health Affairs, through the Joint Program Committee-1/Medical Simulation and Information Science Research Program under Award No. W81XWH-16-1-0460. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense. The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office.

References

1. Pew Research Center: Social Trends, http://www.pewsocialtrends.org/2016/10/06/the-state-of-american-jobs/ (2016)
2. Brynjolfsson, E., McAfee, A.: Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy. Digital Frontier Press, Lexington, MA (2011)
3. Woolf, B. P.: Building Intelligent Interactive Tutors: Student-centered Strategies for Revolutionizing e-Learning: Morgan Kaufman (2008)
4. National Academy of Engineering: Grand Challenges for Engineering. National Academy of Sciences/National Academy of Engineering, Washington, DC (2008)
5. Craig, S. D., Graesser, A. C., Sullins, J., Gholson, B.: Affect and Learning: An Exploratory Look into the Role of Affect in Learning with AutoTutor. Journal of Educational Media. 29(3), 241-250 (2004)
6. Hatano, G., Inagaki, K.: Two Courses of Expertise. In: Stevenson, H., Asuma, H., Hakauta, K. (eds.) Child Development and Education in Japan, pp. 262-272. Freeman, San Francisco (1986)
7. Bransford, J. D., Schwartz, D. L.: Rethinking Transfer: A Simple Proposal With Multiple Implications. In: Iran-Nejad, A., Pearson, P.D. (eds.) Review of Research in Education, vol. 24. American Educational Research Association, Washington, DC (1999)

8. Coultas, C.W., Grossman, R., Salas, E.: Design, Delivery, Evaluation, and Transfer of Training Systems. In: Salvendy, G. (eds). Handbook of Human Factors and Ergonomics, 4th ed, pp. 490-533. John Wiley & Sonss, Hoboken, New Jersey (2012)
9. Pan, S. J., Yang, Q.: A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359 (2010)
10. McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., Scalese, R. J.: A Critical Review of Simulation-based Medical Education Research: 2003-2009. Medical Education. 44(1), 50-63 (2010)
11. Kindley, R. W.: Scenario-based e-Learning: A Step Beyond Traditional e-Learning. Learning Circuits, http://www.learningcircuits.org (2002)
12. Lave, J., Wenger, E.: Situated Learning: Legitimate Peripheral Participation. Cambridge University Press (1991)
13. Studnek, J. R., Fernandez, A. R., Shimberg, B., Garifo, M., Correll, M.: The Association Between Emergency Medical Services Field Performance Assessed by High-fidelity Simulation and the Cognitive Knowledge of Practicing Paramedics. Academic Emergency Medicine. 18(11), 1177-1185 (2011)
14. Vidani, A. C., Chittaro, L., Carchietti, E.: Assessing Nurses' Acceptance of a Serious Game for Emergency Medical Services. In: 2nd IEEE International Symposium on Games and Virtual Worlds for Serious Applications (VS-GAMES), pp. 101-108. IEEE Press, New York (2010)
15. Durlach, P., Spain, R.: Framework for Instructional Technology. In: Duffy, V. G. (ed.) Advances in Applied Human Modeling and Simulation. CRC Press (2012)
16. Pea, R. D.: The Social and Technological Dimensions of Scaffolding and Related Theoretical Concepts for Learning, Education, and Human Activity. The Journal of the Learning Sciences. 13(3), 423-451 (2004)
17. Murray, T., Arroyo, I.: Toward Measuring and Maintaining the Zone of Proximal Development in Adaptive Instructional Systems. In: Proceedings of the 6th International Conference on Intelligent Tutoring Systems (2002)
18. Vygotsky, L. S.: Mind and Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA (1978)
19. Anderson, J. A., Corbett, A. T., Koedinger, K., Pelletier, R.: Cognitive Tutors: Lessons Learned. The Journal of the Learning Sciences. 4(2), 167-207 (1995)
20. Dillenbourg, P., Self, J.: A Framework for Learner Modeling. Interactive Learning Environments. 2(2), 111-137 (1992)
21. Pardos, Z. A., Heffernan, N. T., Anderson, B., Heffernan, C. L.: Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks. In Christobal, R. (ed.) Handbook of Educational Data Mining, pp. 417-426. CRC Press (2010)
22. Vesin, B., Klašnja-Milićević, A., Ivanović, M., & Budimac, Z. (2013). Applying Recommender Systems and Adaptive Hypermedia for e-Learning Personalization. Computing and Informatics, 32(3), 629-659.
23. Dziuban, C., Moskal, P., Johnson, C., & Evans, D.: Adaptive Learning: A Tale of Two Contexts. Current Issues in Emerging eLearning, 4(1), 3 (2017)
24. Buckley, R., Caple, J.: The Theory and Practice of Training. Kogan Page Publishers (2009)
25. United States Department of Transportation, & National Highway Traffic Safety Administration. EMT-Basic: National Standard Curriculum (1996)
26. Sottilare, R. A.: Adaptive Intelligent Tutoring System (ITS) Research in Support of the Army Learning Model: Research Outline. US Army Research Laboratory (ARL-SR-0284) (2013)
27. Duffy, E.: The Psychological Significance of the Concept of “Arousal" or “Activation”. Psychological Review. 64(5), 265 (1957)
28. Berlyne, D. E.: Curiosity and Learning. Motivation and Emotion. 2(2), 97-175 (1978)

29. Cohn, J. V., Kruse, A., Stripling, R.: Investigating the Transition from Novice to Expert in a Virtual Training Environment Using Neuro-Cognitive Measures. In: Schmorrow, D. (ed.) Foundations of Augmented Cognition. LEA, Mattawan, NJ (2005)
30. Cohn, J. V., Nicholson, D., Schmorrow, D. (eds.): The PSI Handbook of Virtual Environments for Training and Education, vol. 3. Praeger Security International, Westport, CT (2008)
31. Kirkpatrick, D. L.: Implementing the Four Levels: A Practical Guide for Effective Evaluation of Training Programs. Berrett-Koehler, San Francisco (2007)
32. Williams, B., Boyle, M., Molloy, A., Brightwell, R., Munro, G.: Undergraduate Paramedic Students' Attitudes to E-Learning: Findings from Five University Programs. Research in Learning Technology. 19(2), 89-100 (2011)
33. Freeman, J. B., Ambady, N.: MouseTracker: Software for Studying Real-time Mental Processing Using a Computer Mouse-Tracking Method. Behavior Research Methods. 42(1), 226-241 (2010)
34. Kawatsu, C., Hubal, R., Marinier, R.: Predicting Students' Decisions in a Training Simulation: A Novel Application of TrueSkill™. IEEE Transactions on Computational Intelligence and AI in Games, pp. 99 (2016)
35. Folsom-Kovarik, J. T.: Developing a Pattern Recognition Structure to Tailor Mid-Lesson Feedback. In Sottilare, R. (ed.) In: Proceedings of the 5th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym5), self published (2017)
36. Wearne, A., Wray, R. E.: Exploration of Behavior Markers to Support Adaptive Learning Lecture Notes in Computer Science. In: Proceedings of the 2018 Human Computer Interaction International (HCII) Conference, Las Vegas (2018)
37. Wray, R.E., Stowers, K.: Interactions Between Learner Assessment and Content Requirements: A Verification Approach. In: Andre, T. (ed.) Advances in Human Factors in Training, Education, and Learning Sciences, pp. 36-45. Springer, Cham, Switzerland (2018)
38. Shute, V. J.: Focus on Formative Feedback. Review of Educational Research. 78(1), 153-189. (2008)

APPENDIX E: Aggregate Results from Assessment of the Role of Behavioral Markers in Adaptive Learning

The following summarizes the statistical analysis of pre- and post-test scores obtained from the study. The analysis was conducted using SPSS version 24.

The following analyses were conducted:

- A paired-sample t-test to assess the general difference in scores between pre and post-data
- A one-way ANOVA with post-hoc (Tukey) test to assess differences between conditions

Assumptions

Assumptions required for each test, as well as the outcomes, are detailed below.

T-test Assumptions

Dependent variable is continuous

- Yes, all measurement was based on test scores, which are recorded on a continuous scale

Independence of observations

- Independence can be reasonably assumed, as the data collection process was random

Normal distribution of DV

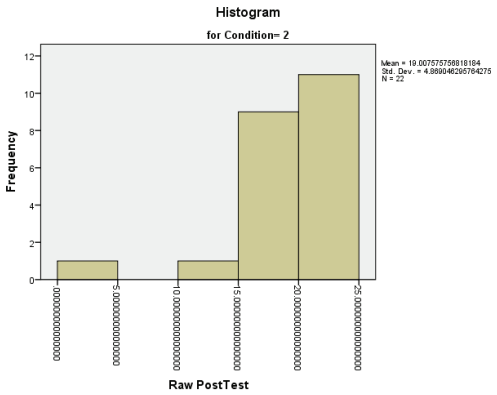
- Since the sample size was larger than 50, the Kolmogorov-Smirnov test was used to assess normality. Output shows a p value of less than .05 for data in condition 2 post-test, suggesting abnormality in the distribution of data (see below).
- Examination of pre-test scores
    - Further inspection of the data shows relative normality for pre-test scores.
- Examination of post-test scores
    - Further inspection of the data shows issues with normality for the post-test scores, specifically in condition 2 (Histogram & Q-Q plot copied below for convenience)

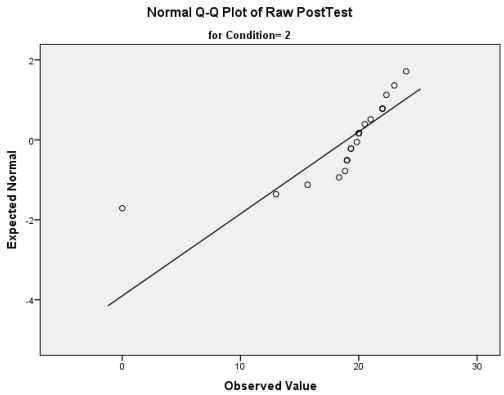
Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Condition	Statistic	df	Sig.	Statistic	df	Sig.
Raw PreTest	1	.111	22	.200*	.958	22	.451
	2	.133	22	.200*	.961	22	.511
	3	.135	18	.200*	.931	18	.202
Raw PostTest	1	.197	22	.026	.948	22	.286
	2	.309	22	.000	.658	22	.000
	3	.137	18	.200*	.950	18	.431

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction





Lack of significant outliers

- Examination of a box-and-whisker plot suggests there may be one outlier (case #25). This case had an unusually low pre-test score of 9 (next lowest score was 10.8). At this point, judgment to remove the data point was withheld, pending further inspection of the data.
- Examination of box-and-whisker plot suggests two outliers, a minor one (case #11) and a major one (case #25). Case #11 had an unusually low score of 13, with the next lowest score above that being 15. We chose not to eliminate this data point, as no additional evidence exists for it being problematic (no evidence of participant lack of participation or other pollution of data). In contrast, case #25 had a post-test score of 0, which is reasonably assumed to be indicative of participant non-participation or experiment session failure. We decided to remove this data point.

ANOVA Assumptions

The following assumptions are covered above:

- Continuous measurement (interval or ratio)
- Independence of observations
- Lack of significant outliers
- Normal distribution

IV must be 2 or more independent groups

- 3 conditions were implemented in this study (between-subjects), resulting in 3 groups

Homogeneity of variance

- Completion of Levene's test for homogeneity of variance showed no significance, thus homogeneity is assumed (see table below)

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
Raw PreTest	1.250	2	58	.294
Raw PostTest	.350	2	58	.706

Results

All output from SPSS analyses are documented in a logfile ("OuputAug2018") that was generated during data analysis and can be retrieved for further reference. As above, prior to analysis, we removed case #25 (post-test score was 0).

T-test Results

A paired-sample t-test was completed to assess the difference in scores between pre and post-data. Based on the analyses, there is evidence ($t = 13.866$, $p < .001$) that the tutoring intervention as a whole improves learning outcomes. This finding does not take into account differences between tutoring strategies implemented. However, it does suggest that the tutoring program as a whole was effective for helping novice learners grow in their knowledge of EMT practices, specifically scene size-up. Results from an ANOVA (next section) can indicate whether a particular method of tutoring was more effective.

SPSS-generated results are included below for reference.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Raw PostTest	19.668	61	2.332	.298
	Raw PreTest	15.371	61	2.452	.314

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Raw PostTest & Raw PreTest	61	.489	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Raw PostTest - Raw PreTest	4.296	2.420	.309	3.676	4.916	13.866	60	.000

ANOVA Results

A one-way ANOVA with post-hoc (Tukey) test was completed to assess differences between conditions. This analysis was completed on both pre- and post-test scores, though the intention was to examine differences between post-test scores between conditions. The analysis included a calculation of partial η^2, an acceptable effect sized measurement for univariate / one-way analyses of variance.

There was a significant difference detected between groups for post-test scores ($F(2, 60) = 3.63$, $p < .05$, partial $\eta^2 = .11$). A Tukey post-hoc test revealed a significant difference ($p < .05$) between scores in condition 1 (M = 15.34, SD = 2.52, 95% CI [14.22, 16.45]) and condition 3 (M = 15.70, SD = 2.82, 95% CI[14.29, 17.10]). In particular, the mean difference between scores in these conditions was 1.86 (95% CI [.15, 3.57]), with participants scoring on average higher in condition 3. This suggests that, for our study, adapting tutoring strategy based on both performance and behavioral markers (mouse movements) was more effective than not adapting tutoring at all. However, it is unclear if adapting tutoring based on performance alone was helpful (this condition was not significantly different from other conditions). Aside from the indication of significance in difference between non-adaptive and fully adaptive (based on performance + markers), the effect size was also medium in nature, giving some additional support for this finding.

SPSS-generated results are copied below for convenience.

Descriptives

						95% Confidence Interval for Mean			
		N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
Raw PreTest	1	22	15.337	2.516	.536	14.221	16.452	10.833	20.666
	2	21	15.126	2.117	.462	14.162	16.091	10.833	20.000
	3	18	15.699	2.818	.664	14.297	17.100	12.250	22.000
	Total	61	15.371	2.452	.314	14.743	15.999	10.833	22.000
Raw PostTest	1	22	18.700	1.960	.418	17.831	19.570	15.00	23.000
	2	21	19.912	2.443	.533	18.800	21.024	13.00	24.000
	3	18	20.564	2.300	.542	19.420	21.708	15.25	24.000
	Total	61	19.668	2.332	.298	19.070	20.265	13.000	24.000

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Raw PreTest	Between Groups	3.213	2	1.607	.260	.772
	Within Groups	357.753	58	6.168		
	Total	360.966	60			
Raw PostTest	Between Groups	36.317	2	18.158	3.631	.033
	Within Groups	290.037	58	5.001		
	Total	326.354	60			

Multiple Comparisons

Tukey HSD

Dependent Variable	(I) Condition	(J) Condition	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Raw PreTest	1	2	.210	.757	.959	-1.612	2.032
		3	-.361	.789	.891	-2.260	1.536
	2	1	-.210	.757	.959	-2.032	1.612
		3	-.572	.797	.754	-2.490	1.346
	3	1	.3619	.789	.891	-1.536	2.260
		2	.572	.797	.754	-1.346	2.490
Raw PostTest	1	2	-1.211	.682	.186	-2.852	.429
		3	-1.864*	.710	.030	-3.573	-.154
	2	1	1.211	.682	.186	-.429	2.85
		3	-.652	.718	.638	-2.379	1.075
	3	1	1.864*	.710	.030	.154	3.573
		2	.652	.718	.638	-1.075	2.379

*. The mean difference is significant at the 0.05 level.