Simulated Tutors in Interactive Learning Environments: Empirically-Derived Design Principles

Natalie B. Steinhauser
Jardon & Howard Technologies, Inc.
Orlando, FL
Natalie.b.martin.ctr@navy.mil

Leslie A. Butler
Psychology Dept. FAU
Boca Raton, FL
lab1573@aol.com

Gwendolyn E. Campbell
Naval Air Warfare Center
Training Systems Division
Orlando, FL
Gwendolyn.campbell@navy.mil

ABSTRACT

The technical capability to design interactive and adaptive learning environments is maturing rapidly. In this paper, we report on an empirical investigation that used a variant of the wizard-of-oz methodology to identify guidelines for the design of effective simulated tutors within this type of environment. Thirty-six subjects completed a 4 hour curriculum on simple direct current (D.C.) circuits while interacting with an unseen human tutor via a chat interface. Student learning, calculated using pre-test and post-test scores, were significantly correlated with the amount that the students talked, as measured by the average number of words per student turn. Also, the use of open-ended questions resulted in a statistically significant increase in the length of student utterances. Tutor strategies that were significantly positively predictive of student learning gains included reinforcing correct responses by restating the answer and remediating incorrect responses by proposing activities likely to help the student figure out the correct answer by him or herself. We suggest that the strategies that proved effective for human tutors are good candidates for incorporation into a simulated tutor. The paper ends with a discussion of the implications for the design of simulated tutors, the limitations of the current research, and additional research requirements.

ABOUT THE AUTHORS

Natalie B. Steinhauser is a contractor for Jardon & Howard Technologies Inc. and is working as a Psychological Research Associate at Naval Air Warfare Center Training Systems Division. She holds a B.S. in Human Factors/Applied Psychology from Embry-Riddle Aeronautical University and is enrolled in the graduate program in Modeling and Simulation at the University of Central Florida. Her research interests include human-computer interactions, training effectiveness, and display layout and design.

Leslie A. Butler is a contractor for Keagan Corp. and is working as a Psychological Research Associate at Naval Air Warfare Center Training Systems Division. She holds a B.A. in Psychology and is enrolled in the graduate program in General Experimental Psychology at Florida Atlantic University. Her research interests include various topics in cognitive psychology including memory processes, attention, and decision making.

Gwendolyn E. Campbell is a Senior Research Psychologist at Naval Air Warfare Center Training Systems Division. She holds an M.S. and Ph.D. in Experimental Psychology from the University of South Florida and a B.A. in Mathematics from Youngstown State University. Her research interests include the application of human performance modeling techniques within training systems and the development of a cognitively based science of instruction.

Simulated Tutors in Immersive Learning Environments: Empirically-Derived Design Principles

Natalie B. Steinhauser
Jardon & Howard Technologies, Inc.
Orlando, FL
Natalie.B.Martin.ctr@navy.mil

Leslie A. Butler
Psychology Dept. FAU
Boca Raton, FL
lab1573@aol.com

Gwendolyn E. Campbell
Naval Air Warfare Center
Training Systems Division
Orlando, FL
Gwendolyn.Campbell@navy.mil

INTRODUCTION

The technical capability to design interactive and adaptive learning environments is maturing rapidly. Not surprisingly, the U.S. military, one of the largest consumers of training and a driving force behind the training industry, wants to leverage technical capabilities in simulation and e-learning to prepare the 21st Century Warfighter (Harris, 2002). The problem is that raw technical capability alone is not sufficient to guarantee optimal training outcomes (Kasvi, 1997). We also must know how and when to utilize those capabilities. The capability to build a simulated tutor, for example, is here. We have tools to design and control the tutor's physical representation (e.g., Allbeck & Badley, 2002), and tools capable of representing domain knowledge and pedagogical procedures (e.g., Weiland et al., 2002). However, despite a large body of research literature investigating human tutoring (e.g., Chi, Siler, Jeong, Yamauchi & Hausmann, 2001; Graesser & Person, 1994; Graesser, Person & Magliano, 1995; McArthur, Stasz & Zmuidzinas, 1990; VanLehn et al., in press; VanLehn, Siler, Murray, Yamauchi & Bagget, 2003), many questions about how a simulated tutor should be designed to behave remain unanswered.

In this paper, we present the results of an empirical study conducted with human tutors, designed to help determine the characteristics of effective tutoring in a conceptual domain. In an attempt to establish some similarity between our experimental context and the context that trainees would likely experience with a simulated tutor, our students and tutors were physically separated and communicated via a (text-based) chat interface. The selected domain was circuits and electricity. While there are obviously many quantitative relationships in this domain, our lessons focused on teaching the concepts and relationships qualitatively.

One of the most obvious differences between tutoring and typical classroom instruction is the nature of the dialogue between the student and the instructor. More specifically, students typically contribute a significantly larger proportion of the dialogue during a tutoring session than during a classroom session. Given the technical difficulties in designing a system that can interpret and generate natural language, our first research question was whether or not eliciting natural language from a student actually has an instructional benefit. Beyond that, we were interested in identifying effective pedagogical strategies used by human tutors and specifying those strategies at a level of description that could be implemented in a simulated tutor.

METHOD

Participants

There were thirty-six participants in the role of student, 19 female and 17 male, recruited from a large southeastern university with a mean age of 22.2 years (SD=4.8). There were three participants in the role of tutor, 2 male and 1 female. All of the tutors had prior experience tutoring and were knowledgeable in the area of electricity and circuits. Two of the tutors had experience providing military technical training. Student and tutor participants received monetary compensation for their time.

Apparatus

Two Dell Latitude laptop computers were set up in a single room, visually separated by a room divider, and networked via an Ethernet hub and cables. The student sat at one workstation and the tutor sat at the other. In addition to the laptop, the tutor's workspace also had a second monitor, which was synched to the student's computer and allowed the tutor to observe everything that the student was doing. The network connection allowed the two participants to communicate in real time with a chat interface, in lieu of speaking to one another.

Materials

A learning environment was developed to teach a set of basic concepts and procedures in electricity and circuits. As can be seen in Figure 1, this environment presented the student with three separate display areas. One section served as a page turner with standard navigation capability, presenting didactic information and posing problems and questions to the student through a series of lesson slides. The second area was a simulation-based circuits workbench that allowed students to build, load, manipulate and measure any circuit (series or parallel) that could be created with up to ten light bulbs, five batteries and five switches. The majority of the problems posed by the lesson slides required that the student conduct some sort of activity in this simulation environment. Finally, the third area provided the student with a chat interface for communicating with the tutor. Students were instructed to supply their answers to all of the questions posed by the lesson slides to the tutor via this interface. In addition, they were encouraged to communicate with the tutor at any time, if they had questions or needed help progressing through the lesson. The tutors were encouraged to "talk" with their students in whatever manner they felt would be most effective, within the parameters of the established curriculum. For example, they were asked not to change the presentation order of the topics and to be consistent with the vocabulary presented in the lesson slides. The system saved the transcripts from these chats, as well as all of the students' keystrokes in the learning environment.

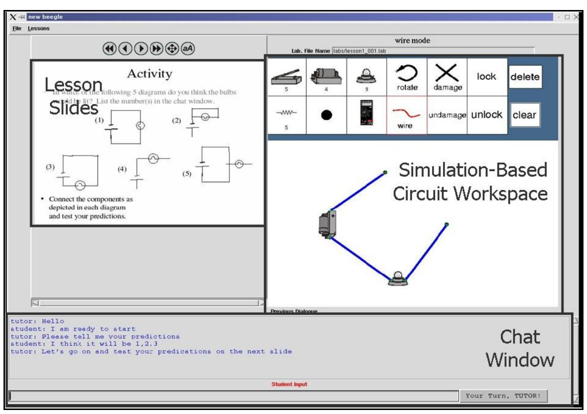

Figure 1. Screen shot of student's learning environment.

The learning environment runs in SUSE Linux. The Graphical User Interface (GUI) was programmed in Tcl/Tk (www.tcl.tk). In order to generate valid voltage and current values for the circuits, the workbench GUI was connected to the gnucap circuit simulator agent (www.gnu.org/software/gnucap/) via the Open Agent Architecture (OAA) (www.ai.sri.com/~oaa/).

Two lessons were created as series of slides in Microsoft PowerPoint. The first lesson addressed basic circuit components, complete and incomplete circuits and fault detection by measuring voltage, all within the context of series circuits. The second lesson expanded on these topics within the context of parallel circuits. The bulk of the curriculum was presented in the following sequence: After the didactic presentation of a minimal amount of information (e.g., the introduction of a new component, symbol or concept), students were asked to predict some aspect of the behavior of a circuit and explain the basis for their prediction. Then students were asked to build (or load) that circuit and test their predictions by some combination of observation and/or measurement. Finally, students were asked a series of discussion questions to compare the actual results to their predictions and to draw general conclusions about properties and principles of circuits.

In order to support our investigation of the role that unconstrained language plays in determining the effectiveness of a training experience, three versions of these lessons were created that varied only in the format of the questions that the students were asked. The initial version included open-ended questions, such as "What do you think will happen?" and "Why?" whenever this was appropriate. In contrast, a closed-ended version was created that solely used multiple choice questions. This was often done by asking the same question as in the open-ended version and then describing (in a multiple choice format) different possible answers suggested by different hypothetical students. The participants were asked to select the hypothetical student that they most agreed with. Finally, we attempted to create a compromise, or "middle-of-the-road," version of the lessons that used short answer questions. It often required two or three short answer questions in a row to elicit the same information that was associated with a single open-ended "Why?" question.

Two tests, with multiple-choice questions covering the topics from the lessons, were also created. The pre-test had 38 questions, and the post-test had 21 questions. Finally, we wrote a short demographics questionnaire that asked about the participant's background in circuits and electricity, and a short questionnaire asking the participant to evaluate the lesson materials and tutor along several dimensions, such as understandability and helpfulness.

Design

There were two independent variables. The first was the tutor. Participants were semi-randomly assigned to one of our three tutors, such that each tutor worked with twelve participants (perfect random assignment was not possible due to the tutors' work schedules). The second independent variable was lesson condition: open-ended, middle-of-the-road and closed-ended. Participants were also randomly assigned to a lesson condition, such that each tutor ran four participants through each lesson condition.

Procedure

The experimental session lasted approximately 4 hours. Participants were run individually. After completing the informed consent paperwork and filling out a short demographic survey, participants took the pre-test and were shown how to work in the computer-based training environment. Then they worked through the first lesson with a tutor. After completing the first lesson, they were asked to fill out a short questionnaire evaluating their experience and offered a break. Next, they completed the second lesson with the same tutor, filled out the same questionnaire again (addressing the second lesson) and were offered another break. Finally, they took the post-test and were debriefed.

RESULTS

Data from four participants had to be discarded due to technical difficulties with the learning environment during their experimental sessions. Results from the remaining 32 participants are presented below.

Role of Student Language

Our system was developed based on the hypothesis that participating in a tutorial dialogue would be an effective way for a student to learn. In order to evaluate this hypothesis, we quantified student participation by calculating the average number of words per turn in each student's dialogue. For this analysis, we used change scores (post-test score – pre-test score) to assess student learning. As hypothesized, the average number of words per turn in the students' dialogue was significantly positively correlated with their change scores, $r\ (32) = 0.396$, $p = .025$.

Next, we ran a two-way analysis of variance (ANOVA) to determine whether or not asking open-ended questions encouraged more dialogue from the student (the other independent variable was the tutor, and it did not return a significant effect, nor a significant interaction). The results were significant, $F\ (2,25) = 11.77$, $p < .005$. Follow-up least-significant-difference (LSD) tests revealed that the use of open-ended questions resulted in student responses that were significantly longer (as measured by average number of words per turn) than asking them short-answer or multiple choice questions. Table 1 shows the means and standard deviations for the three lesson conditions.

Table 1. Means and standard deviations of the average number of words per turn in student dialogue based on type of question.

Condition	Mean	Standard Deviation
Open-ended	7.52	2.2
Middle-of-the-road	5.11	1.6
Closed-ended	3.80	1.1

Finally, we ran a mixed model ANOVA to look for a direct relationship between type of question and student learning. Testing time was the within-subjects variable, with two levels (pre-test and post-test). Condition and tutor were the two between-subjects variables. There was a significant main effect of testing time, $F\ (1,25) = 317.2$, $p < .01$, with post-test scores ($M = 0.79$, $SD = 0.14$) being statistically significantly higher than pre-test scores ($M = 0.40$, $SD = 0.11$), indicating that our subjects did learn. None of the other main effects or interactions were significant.

Coding Tutor Strategies

Moving beyond the nature of the established curriculum, we were interested in determining what strategies, if any, a tutor could use during the dialogue to yield a positive impact on student learning. We reviewed a small subset of the transcripts (selected for the variety of tutor-student interactions) and culled examples of tutor responses to both correct and incorrect student answers. With input from one of the tutors, we developed a classification system to cover those responses. See Table 2 for a list of those codes and a brief description of each.

Three independent coders annotated a subset of the dialogue transcripts. Following Carletta et al. (1997), we calculated kappa as our pair-wise measure of inter-rater reliability. Our coders achieved a "near perfect" average kappa of 0.91 when distinguishing, in the tutor dialogue, between reinforcement and remediation segments, and "substantial" average kappas of 0.70 and 0.78 when distinguishing amongst the types of reinforcement strategies and the types of remediation

Table 2. List of categories used to code common tutor strategies found in our corpus of dialogue transcripts, with descriptions.

Strategy Classification	Strategy Name	Description
Reinforcement		Types of responses used when student's answer is basically correct.
	Accept and Move on	Tutor acknowledges answer, indicates its acceptability and signals the student to continue (either implicitly or explicitly). No domain content is included.
	Restatement	In addition to acknowledging the answer and its acceptability, the tutor restates some domain content. The restatement may serve to reiterate a key point, to summarize across several turns, to model more appropriate use of technical terminology or simply to express the same ideas in different words.
	Extend	In addition to acknowledging the answer and its acceptability, the tutor brings in additional domain content that is not the focus of the current segment of the lesson. The tutor may bring up related but new concepts that will be addressed later in the curriculum, or mention related real world experiences and/or situations.
Remediation		Types of responses used when student's answer is basically incorrect.
	Bottom Out	Tutor provides correct answer to student.
	General Prompt	Tutor prompts student to elaborate or revise answer, e.g., "Are you sure?" and "Is that all?"
	Restate Question	Tutor tries to elicit correct answer by restating the question, often trying to make it easier by making it more concrete or providing choices.
	Give Hint	Tutor tries to help student produce the correct answer by giving a hint of some type.
	Activity-based	Tutor recommends activity that should yield correct answer for student. This activity may involve re-reading earlier lesson material or conducting a test in the simulation workspace (building or measuring a circuit).

strategies respectively. Once it was established that our coders could annotate tutor strategies reliably with this scheme, the remainder of the transcripts were coded.

After the tutor strategies in each transcript were coded, we counted the number of instances of each strategy type used when interacting with a particular student and transformed those counts into percentages. So, for example, we could see that, for one student, 75% of the reinforcements he received were of the type "accept and move on" and only 25% were of the type "restatement", while for another student, 40% of her reinforcements were "accept and move on" and 60% were "restatements".

Evaluating Tutor Strategies

Finally, in order to evaluate the effectiveness of the various reinforcement and remediation strategies, we ran multiple linear regression analyses, attempting to predict student post-test scores with the extent to which they experienced various tutor strategies. We conducted separate analyses for reinforcement and remediation strategies. We also used the students' pre-test scores and self-reported SAT scores as predictors. One final predictor used in each analysis was the frequency with which students were either reinforced or remediated during the lessons. Thus, for example, the reinforcement analysis poses the question, if one controls for incoming knowledge, general academic ability and the frequency with which a student received any reinforcement at all, does the frequency with which the student received a particular reinforcement strategy predict how well the student did on the post test?

Tutor Strategies When Student is Correct

First, consider reinforcement strategies. Using a stepwise method, a significant model emerged: $F(3,26) = 17.972$, $p < .005$. The model explains 66.2% of the variance in the post test scores (Adjusted $R^2 = 0.662$). Table 3 gives information for the predictor variables that are included in the model. As can be seen in the table, one particular method of reinforcement stands out as positively predictive of post-test scores, restatements of the correct answer. It can also be seen that the extent to which a student is reinforced during the tutorial session is a significant positive predictor of post-test scores, even after the variance associated with more incoming knowledge is removed. In other words, after controlling for incoming knowledge, students who complete the lessons easily, with few mistakes, do best on the post-test.

Table 3. The unstandardized and standardized regression coefficients for the variables included in the reinforcement model.

Variable	B	SE B	Beta
Pre-test score	.332	.113	.339*
Reinforcement	.917	.143	.885*
Restatement	.551	.172	.442*

* $p < .005$

Tutor Strategies When Student is Incorrect

Next, consider remediation strategies. Again, using a stepwise method, a significant model emerged: $F(3,27) = 19.305$, $p < .005$. The model explains 67% of the variance in the post test scores (Adjusted $R^2 = 0.670$). Table 4 gives information for the predictor variables that are included in the model. As can be seen in the table, one particular method of remediation stands out as positively predictive of post-test scores, suggesting an activity that should help the student figure out the correct answer. It can also be seen that, as expected in light of the previous results, after controlling for incoming knowledge, students who have difficulty completing the lesson (and thus receive a large amount of remediation) also have difficulty on the post-test.

Table 4. The unstandardized and standardized regression coefficients for the variables included in the remediation model.

Variable	B	SE B	Beta
Pre-test score	.277	.134	.239*
Remediation	-.921	.142	-.789*
Activity-based	.585	.232	.302*

* $p < .05$

DISCUSSION

This study was designed to generate guidance for the development of a simulated tutor in a computer-based learning environment. We first attempted to determine the importance of the capability of the system to support a natural language dialogue between the student and the tutor. Our results show that the verbosity of the student (as measured by the average number of words per student turn) did correlate positively and significantly with the amount the student learned. While 0.40 is not a particularly large correlation, it should be noted that this is based on pure volume of student language, and doesn't take into account the quality of that language. In other words, just getting the students to talk at all is useful. We also found that it is relatively easy to elicit more verbose utterances from the students by simply asking open-ended questions.

Given these two results, it is perhaps somewhat puzzling that, in our study, students in the open-ended condition did not show statistically significantly more learning gains than students in the other conditions. We hypothesize that this is because our conditions were not truly "pure". The lesson slides certainly followed our condition specifications; however, our participants didn't always "cooperate". Some students generated lengthy answers to multiple choice questions, while other students were brief even when asked open ended questions. In addition, even though our tutors were encouraged to restrict their students' language in the closed-ended condition, when students in this condition had difficulty, it was not uncommon for their tutors to engage them in verbose dialogues. Thus, a true empirical evaluation of this question could best be conducted with a simulated tutor that would behave according to the designated rule set.

Our research has also found correlational support for the implementation of two tutor strategies. First, when students answer questions correctly, our data suggest that restating the correct answer may be more effective than simply acknowledging it and allowing the student to move on. Why might this be helpful? Looking across our data we can identify several pedagogical roles that the reinforcements seemed designed to serve. In some cases, the tutor appears to use the restatement to model a better use of technical terminology (e.g., "That's right. We say that the circuit must be *complete*."). In other cases, the restatement draws the student's attention to the most important bit of information in the answer (e.g., "Exactly! The bottom line is that it doesn't matter *where* the gap is.") We also saw cases where the tutor used a restatement to summarize information that had previously been distributed across multiple turns in the dialogue. It is not difficult to imagine that all of these types of restatements could be helpful to a student becoming familiar with a new domain.

Next, when students answer questions incorrectly, our data suggest that proposing an activity that should help the student figure out the answer may be the most effective strategy. The "why?" behind this result is fairly obvious. Most cognitive theories of learning propose that learning, especially meaningful learning (as compared to rote learning) is a function of the amount of cognitive effort and processing that a student applies. The other remediation strategies found in our corpus – simplifying the question, giving a hint, giving the answer – all tend to markedly reduce the requirement on the student. Suggesting an activity that might help resolve the problem is a way to provide guidance without necessarily reducing the student's processing burden in any significant way.

To summarize, this study suggests three guidelines for the development of simulated tutors in computer-based training systems. First, the systems should encourage the students to "talk" about the concepts being taught. We found that the incorporation of open-ended questions is one way to elicit more language from students. Second, simulated tutors should be able to reinforce correct student answers by restating the concepts and relationships in those answers. Finally, simulated tutors should be able to remediate incorrect student answers by proposing activities designed to help students figure out the correct answer.

Of course, there are at least two limitations to this research that should be considered. First, our data are correlational, and thus drawing causal inferences is a risky business. It is difficult to determine if it was the tutor's strategy that led to the student's learning gain, or if it was something else entirely – possibly some characteristic of the student – that is responsible for the learning gain accrued and also evoked a particular strategy usage from the tutor. For example, maybe it was the student's natural aptitude for the material that determined the learning gain, and tutors instinctively only used the "activity-based remediation" strategy with the brightest students, because they felt that students with less aptitude would not be able to successfully complete the recommended activity and draw the correct conclusions from it.

Second, generalizing from a human tutor to a simulated tutor is also a bit risky. It is possible that students expect (and/or are willing to tolerate) different types of behavior from another human than from a computer. While there are some indicators suggesting that people will go to great lengths to ensure that their interactions with a computer are as human-like as possible (e.g., Weizenbaum, 1976), there is also evidence that people interact with computer tutors differently than with human tutors (e.g., Litman & Forbes-Riley, 2006). Thus, we don't know for sure that, even if these strategies were directly responsible for some amount of learning gain in our students, they are the very best ones to implement in a simulated tutor.

Ultimately, research on the design of an effective simulated tutor should be conducted with a simulated tutor. The fact that simulated tutors will follow rules perfectly allows researchers to set up experiments that can yield causal data. In addition, of course, using simulated tutors in your research circumvents any concerns about generalizing from human to simulated tutors. However, until these simulated tutors exist that are robust enough to support training and training research, empirical examinations of human tutor behavior and its effectiveness seem like the most reasoned and reasonable place to start our search for preliminary design guidance. Technological advances will keep coming at a breakneck pace, regardless of whether or not we provide any empirically-based guidelines for their development.

Our warfighters face increasing requirements to deal with a variety of joint organizations and missions, and to respond effectively to changing threats and asymmetric enemies. Training has always been one of the military's most powerful tools to prepare our warfighters. The advances in training technologies should be able to meet our current and future challenges – but only if we design them intelligently, based on empirically derived principles and guidelines.

ACKNOWLEDGEMENTS

We would like to thank our sponsors from the Office of Naval Research, Dr. Susan Chipman and Dr. Ray Perez; our collaborators at the University of Edinburgh, Dr. Charles Callaway, Dr. Myrosia Dzikovska, Elaine Farrow and Professor Johanna Moore; two former Research Associates who worked on this project, Lisa Durrance and Cheryl Johnson.

REFERENCES

Allbeck J. & Badler, N. (2002). Embodied autonomous agents. In K. Stanney's (Ed.) *Handbook of Virtual Environments*. Lawrence Erlabuam Associates: Mahwah, NJ. pp. 313-332.

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G. & Anderson, A. G., (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, *23*(1), 13-31.

Chi, M., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. (2001). Learning From Human Tutoring. *Cognitive Science, 25*, 471-533.

Graesser, A. & Person, N. (1994). Question asking during tutoring. *American Educational Research Journal, 31*(1), 104-137.

Graesser, A., Person, N., & Magliano, J. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*, 495-522.

Harris, P. (2002, November). How the U.S. military is reinventing learning. *Learning Circuits, 3,* n11. Retrieved April 27th, 2007 from http://www.learningcircuits.org/2002/nov2002/harris.html

Kasvi, J. (1997). [Review of the book *Technology-based training: A comprehensive guide to choosing, implementing, managing, and developing new technologies in training*]. Retrieved April 27th, 2007 from http://www.knowledge.hut.fi/projects/itss/referRavet.pdf

Litman, D & Forbes-Riley, K. (2006). Correlations between dialogue acts and learning in spoken tutoring dialogues. *Natural Language Engineering*, *12* (2), 161 - 176.

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. *Cognition and Instruction, 7*, 197-244.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (accepted for publication). When are tutorial dialogues more effective than reading? *Cognitive Science*.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*, 209-249.

Weiland, W., Szczepkowski, M., Urban, G., Mitchell, T., Lyons, D. & Soles, R. (2002). Reusing cognitive models: Leveraging SCOTT technology in an LCAC virtual training environment. *Proceedings of the Eleventh Conference on Computer-Generated Forces and Behavior Representation*, 315-322.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* New York, NY: W.H. Freeman & Co.