Improving the Efficiency of Dialogue in Tutoring

Kristopher J. Kopp[a,*], M. Anne Britt[a], Keith Millis[a], and Arthur C. Graesser[b]

[a]Department of Psychology

Northern Illinois University

DeKalb, IL, 60115, USA

[b]Department of Psychology

University of Memphis

Memphis, TN, 38152, USA

*Corresponding Author:
Kristopher J. Kopp
Department of Psychology
Northern Illinois University
DeKalb, IL, 60115
Email: kkopp@niu.edu
Phone: 815-753-9121
Fax: 815-753-8088

Abstract

The current studies investigated the efficient use of dialogue in intelligent tutoring systems that use natural language interaction. Such dialogues can be relatively time-consuming. This work addresses the question of how much dialogue is needed to produce significant learning gains. In Experiment 1, a full dialogue condition and a read-only control condition were compared with a mixed dialogue condition in which students engaged in full dialogue for half the problems followed by problems requiring only a limited engagement. We found that the mixed dialogue condition produced results as impressive as the full dialogue condition and took less time. Experiment 2 replicated these findings and further examined issues of time engaged in learning, quality of instruction, and learning gains. Overall, these results show that dialogue-based intelligent tutoring systems could be designed in a more efficient manner to maximize learning and minimize the cost of time-on-task.

Keywords: Intelligent Tutoring Systems; Natural Language Dialogue; Efficient Learning; Scientific Reasoning.

Improving the Efficiency of Dialogue in Tutoring

1. Introduction

Intelligent tutoring systems (ITSs) are computer programs that promote learning by interacting with students in ways similar to those of human tutors interacting with their clients. For example, ITSs pose problems and questions, evaluate the quality of responses and, when necessary, give feedback, hints and prompts in an effort to maximize learning. Natural language dialogue is becoming more prominent in the way that ITSs and other learning environments communicate with students (Atkinson, 2002; Author/s, 2008; Author/s , 2009[a]; Baylor & Kim, 2005; Cole et al., 2003; Johnson, Rickel, & Lester, 2000; Lane & VanLehn, 2005; Litman et al., 2006; McNamara, Levinstein, & Boonthum, 2004; Moreno & Mayer, 2007). Dialogue-based ITSs are dynamic and engaging learning environments that enable learners to take an active part in the knowledge building process.

Systems with interactive dialogue show greater learning gains than less interactive instruction when the emphasis is on deeper levels of knowledge construction (Author/s, 2001; Author/s 2004[a]; Author/s, 2007[a]; Author/s, 2008). Deep knowledge acquisition generally requires reasoning of a logical, causal, or goal-oriented nature (Author/s, 1994; Author/s, 2003) and may involve functional knowledge of specific concepts. Alternatively, shallow knowledge refers to definitions and lists of features of concepts, without a coherent conceptual understanding and a foundation for applying the knowledge to new situations. This is not to say that shallow knowledge cannot be acquired through the use of dialogue during a tutoring session. However, the acquisition of deeper knowledge benefits most from a high quality dialogue.

Effective dialogue in ITS environments has been modeled after interactions between human tutors and their clients. Such conversations often follow predictable patterns that include

the elicitation of student contributions through questions and prompts (Author/s, 1995). Such dialogue moves afford students the opportunity to identify misconceptions or knowledge gaps and to repair these problems. Additionally, such moves promote deep learning by providing students with the opportunity to ask and answer questions, to self-generate ideas, and to produce self-explanations of challenging content (Chi, de Leeuw, Chiu, & LaVancher, 1994; Moreno & Mayer, 2007; Renkl, Stark, Gruber, & Mandl, 1998). The amount of this type of dialogue required to produce significant learning gains remains unclear. This is an important question because the time-consuming nature of dialogues is a potential liability to ITSs. Adequately answering a problem in an ITS may require anywhere from 50 to a few hundred turns (Author/s, 2005). This is not a problem to the extent that the dialogue leads to high quality learning through active engagement. However, if the time-on-task fails to add quality, then it is taking up time that could be spent on other material. Consequently, educators may be hesitant to incorporate ITSs in the classroom due to time demands. Classroom instruction may be perceived as being superior to time spent interacting with a computer because more topics can be covered in a class period using classroom instruction. In general, time-on-task is correlated with learning (Fredrick & Walberg, 1980; Karweit & Slavin, 1981; Taraban, Rynearson, & Stalcup, 2001), but this relationship may depend on the quality of the instruction (Carroll, 1963; Carroll & Spearritt, 1967). Therefore, the goal of this work was to assess whether the general benefit of dialogue-based ITSs can be designed to be more efficient.

1.1. AutoTutor

The ITS that we investigated was AutoTutor (Author/s, 2001; Author/s, 2004[a]; Author/s, 2007[a]; Author/s, 2008). The architectural framework of AutoTutor was inspired by research on adaptive tutoring systems (Anderson, Corbett, Koedinger, & Pelletier, 1995; VanLehn et al.,

2002) and investigations of collaborative learning during human tutoring (Author/s, 1994; Author/s, 1995; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Fox, 1993; Moore, 1995; Shah, Evens, Michael, & Rovick, 2002).

AutoTutor engages a student in a *mixed initiative* dialogue with life-like animated pedagogical agents (see Figure 1) that are synchronized with speech (Author/s, 2003, Author/s, 2007[a]). Both the human and the agent involved in the conversation have the opportunity to participate. The problems or questions presented by the agent typically require roughly a paragraph of information to answer correctly (3-7 sentences). A lengthy answer may be required, but a student's initial answers to questions are typically short (1 or 2 sentences) (Author/s, 2004[a]; Author/s, 2005). It is at this juncture that a dialogue is helpful. AutoTutor engages the student in a dialogue by asking questions and giving feedback to answers in an attempt to get the student to produce a correct and complete answer through self-explanation and self-generation.

A *curriculum script* guides the conversation (see Table 1). This script contains a problem statement and a list of *expectations* (i.e., anticipated correct answers). Each expectation expresses one or two propositions essential to a complete answer. The script also contains appropriate hints, prompts, assertions, and answer summaries (see Appendix for an actual student-tutor dialogue). When a student's initial response fails to correspond to the expectations, AutoTutor will pump the student for more information (e.g., "And can you add to that?"). If the appropriate expectations still have not been covered, a series of hints and prompts are provided. A hint is a leading question or a factual statement intended to guide the learner toward a correct answer. A prompt is a question in which the desired answer is a word or a short phrase that completes the expectation that was missing from the student's contribution (see Table 1 for

example hint and prompt). When a correct response is provided that meets the criterion of an expected answer, an assertion is provided by the tutor. The assertion is simply the expectation, presented in a conversational way that expresses confirmation of a correct response. Assertions are also given when the hints and prompts are not sufficient to get the student to articulate the answer. Based on the self-generation principle (Slamecka & Graf, l978) , it is ideal to get the students to express correct answers rather than having the tutor deliver the correct information. When all expectations have been covered for a particular problem or main question, a summary of the presented information is provided

AutoTutor's feedback is based on the control structure of the program and the response of the student to the various questions. The student input is initially matched to the curriculum script to compare the content of answers to the scripted expectations (Author/s, 2005). This matching process is based on methods used in computational linguistics and Latent Semantic Analysis (LSA, Landauer, Foltz, & Laham, 1998). Based on the comparison results, AutoTutor decides what type of feedback to provide to the student. In addition to hints and prompts, it has the ability to provide positive (e.g., "Perfect answer!"), negative (e.g., "No"), and neutral feedback (e.g., "Okay,"). The feedback is quickly delivered as a direct response to the student's contribution and is intended to be formative (Shute, 2008).

Several experiments have shown the effectiveness of AutoTutor in improving learning in the domains of science and technology, with average effect sizes of .8 standard deviations compared to suitable control conditions (Author/s, 2001; Author/s, 2004[a]; Author/s, 2007[a]; Author/s, 2007[b]). These investigations have often compared a full dialogue condition to control conditions that do not involve a tutor-student interaction. There are two types of control conditions that have been used in these studies. In a *text-only* control condition, participants read

a text covering the same information presented during the AutoTutor session. Alternatively, in a *do-nothing* condition, participants engage in a pre and posttest, but are not exposed to any material regarding the concepts tested.

1.2. Efficiency of Learning

Time is an important aspect of the factors that Carroll (1963) identified as influencing potential learning gains in a traditional educational arena. Students need to be given time (opportunity) to learn and they need to be willing to devote time (perseverance) to the learning activity. Across many different contexts, researchers have found a relationship between time spent with material and learning (Fredrick & Walberg, 1980; Karweit, 1984, Karweit & Slavin, 1981; Taraban et al., 2001). A review of the literature regarding efficiency by Karweit (1984) cites correlations ranging from .10 to .70. Significant correlations (.35) of time and learning have also been found in an online tutoring environment (Taraban et al., 2001). Time-on-task seemingly matters but the relationship is not simply that more time leads to more learning. If that were the case, then one would expect larger correlations. Indeed, Carroll & Spearritt (1967) noted that both the quality of instruction and one's ability to learn also play a critical role in successful learning.

The interactions provided by AutoTutor are designed to create student-centered exchanges that help students in their ability to comprehend the content. While the incorporation of dialogue into a tutoring session is beneficial (Author/s, 2001; Author/s 2004[a]; Author/s, 2007[a]; Author/s, 2008), it tends to be time consuming. It is important to consider whether the positive effect of time might asymptote during learning. In the context of ITSs like AutoTutor, one can imagine that there is a point during the tutoring session in which more time spent engaged in a dialogue will become less beneficial. For instance, the student may master the material early but

be unable to end the session. Alternatively, the student may require time to reflect on the material. Or the student may require additional information to comprehend the targeted content. Without knowing what precise information would aid in the student's understanding, giving additional problems or additional dialogue may not be of any help. Lastly, the student may get bored or tired and stop paying attention. In each of these cases, more time spent with extra dialogue would result in a diminished return (Walberg & Tsai, 1984). The law of diminishing returns refers to the relationship between two variables when one variable is incrementally increased and the results of an outcome variable diminish in size (Cobb & Douglas, 1928; Walberg, 1983).

In relation to the current study, the increasing variable would be time and the outcome variable would be learning gains. Time spent with material may increase learning to a certain point, however, after that point is reached less and less learning may occur as a function of per unit of time resulting in a diminished return. Thus, an efficient learning environment would conceivably be one which minimizes time-on-task (opportunity and perseverance) while maximizing learning, and thereby reducing the possibility of a diminishing return.

Efficiency can be thought of as the "ability to reach established goals with a minimal expenditure of time, effort, or cognitive resources" (Hoffman & Schraw, 2010, p 2). One way to consider efficiency is based on the likelihood model (Hoffman & Schraw, 2009; Perfetti, 1985) in which efficiency is interpreted as the ratio of performance (i.e., learning gains) to effort (i.e., time-on-task). Dividing performance scores by effort scores provides a measure of efficiency between the two variables. Similar to the cost-benefit model (Rossi, Lipsey, & Freeman, 2004), computations are referred to as relative gain or a rate-of-change of performance to effort. One way to conceptualize this is if a student spends less effort (i.e., time) and achieved a better

performance score (i.e., learning gains) than another student, then the first student would be considered to have learned more efficiently. Likewise, if a student achieves a higher gain than another, but both students spend the same time-on-task, then the first student would be considered to have learned more efficiently. One caveat to interpreting these efficiency scores is comparing efficiency scores may only be meaningful in an instructional context to the extent that learning gains reach some criterion.

1.3 Current Studies

While prior studies have investigated the effect of tutorial dialogue on learning, very few have assessed the relationship between time and learning from such tutoring systems. In the current experiments, we were interested in empirically testing the relationship between the amount of learning that occurs in conditions with varying amounts of dialogue interactivity, with an eye towards identifying the most efficient version. In doing so, we expected that the amount of time spent with the materials would differ across conditions.

In both experiments, participants were randomly assigned to one of four tutoring conditions. Three of the four conditions required an interaction between the participant and the tutor, with varying amounts of dialogue beyond the problem statement. A fourth condition required no dialogue interaction. Each condition exposed participants to information using the same method of presentation using AutoTutor. A problem statement and the summary of correct answers were presented by animated agents in all conditions. Only the amount of dialogue varied among conditions.

A *full dialogue* condition used the default version of AutoTutor, incorporating hints and prompts (see Appendix for examples). In a *limited dialogue* condition, participants were given one opportunity to correctly answer the question raised by the problem. No hints or prompts

were given, although a summary of the correct answer was given. A *mixed dialogue* condition provided students with the full dialogue interaction for the first half of the problems, whereas the second half was presented in the fashion of the limited dialogue condition. A *no-dialogue* condition required no interaction between the tutor and the participant. Participants in this condition were not required to answer any questions. The animated agents simply read the problem and summary solution to the participant.

The tutoring content was designed to teach scientific inquiry skills in the domain of Psychology research methods. It presented problems involved in correctly conducting science, rather than simply describing the scientific method. Such skills include evaluating components of experimental designs and correctly reporting, analyzing and interpreting data. The participants were presented with the descriptions of six studies and asked to identify any design flaws. The flaws that were present across these six problems were related to the following eight experimental design concepts:

1. Need for comparison groups
2. Variable must address hypothesis
3. Need for appropriately sensitive dependent measures
4. Need to rely on objective scoring
5. Assurance that effects of mortality or attrition are limited
6. Assurance that effects of subject biases are limited
7. Employ an adequate sample size
8. Assurance that effects of experimenter bias are controlled

1.4 Hypotheses and Predictions

The current experiments included three different types of dependent variables: performance, time-on-task, and efficiency. Performance was assessed by considering learning gains. Time-on-task refers to the time that each student interacts with the tutor during the learning session. Efficiency was assessed using calculated efficiency scores based on the likelihood model described in section 1.2.

The first hypothesis, the *Dialogue Interaction Hypothesis* (Author/s, 2007[a]), assumes that tutorial dialogues increase learning due to the active engagement of the learner during knowledge construction (Aleven & Koedinger, 2002; Chi et al., 1994; VanLehn, Jones, & Chi, 1992) and the collaboration that occurs between the learner and the tutor (Author/s, 1994; Author/s, 1995; Chi, et al., 2001; Fox, 1993; Moore, 1995; Shah et al., 2002). This hypothesis, predicts that conditions containing more dialogue in the AutoTutor context should perform better than those with less dialogue (Hypothesis 1). This hypothesis does not address efficiency.

The second hypothesis is a *Hypothesis of Diminishing Returns* (Hypothesis 2). Similar to Hypothesis 1, it is also assumes an increase in learning due to tutorial dialogue. With increased dialogue one would assume that there would be an increase of time. However, there is a nonzero probability that during some point during the tutorial, an increase in the dialogue will not yield an increase in the resulting learning and thereby the efficiency in its impact will decrease. We expect that the full dialogue condition would be the most likely condition in which students would reach the point of diminished returns because of its length and its heavy reliance on dialogue. If this is the case, then the full dialogue condition would be less efficient than the other conditions. However, we only wanted to compare the efficiency of conditions when the quality of the dialogues was comparable. Therefore, we only compared the efficiency of the full dialogue condition and the mixed dialogue condition because they contained the same dialogue

moves except that the latter only had one-half of the dialogues than the former. According to the Hypothesis of Diminishing Returns, a full dialogue condition will be less efficient than a mixed dialogue condition, as measured by efficiency scores.

2. Experiment 1

2.1. Method

2.1.1. Participants. The participants were 138 native English-speaking undergraduate introductory psychology students from a large Midwestern University who participated for course credit. Seven participants did not return for the second experimental session. Two participants were unable to complete the study due to computer failure. The attrition rates by condition were: no dialogue = 2, limited dialogue = 2, mixed dialogue = 1, full dialogue = 4. This left a total of 129 participants. Presumably, the knowledge level of these students regarding the concepts in this experiment was that of a novice because none of these students had taken a research methods course in psychology. Average pretest scores show that participants accurately answered only about 25% of the items.

2.1.2. Materials

2.1.2.1. Tutor Content. This version of AutoTutor uses two animated pedagogical agents to present flawed studies for students to critique (see Figure 1). The agents presented each participant with six problems that described a research study containing methodological flaws relevant to the target concepts (e.g., a study without a control group). While each problem did not contain flaws pertaining to all 8 concepts, the number of flaws in each problem ranged from three to six. Initially, we had intended to examine the effect of frequency (two mentions vs. four mentions of a flaw). Thus, in Experiment 1, frequency was manipulated as a within-subjects variable. This factor showed no significant effects and for simplicity's sake will not be

mentioned hereafter. For the conditions that required a response from the student, a corresponding curriculum script was used to guide the tutoring session (see Table 1).

2.1.2.2. Assessment. Participants' knowledge of the target scientific reasoning concepts was tested with a critique experiment (CE) task that contained problems similar to those presented during the tutoring session. Participants were asked to critique the experimental design of three scientific studies. Across the three problems, each of the eight flaws was present once. Two versions of this test were created (with a total of six problems) and counterbalanced across pre and posttests. The pretest provided an assessment of students' prior knowledge while differences between pre and posttest provided an assessment of how much was learned from the tutoring session. To assess whether or not these two forms were equivalent assessments, pretest scores were analyzed to identify any differences.

2.1.3. Procedure. The first session consisted of a pretest CE assessment followed by a tutoring session. Participants were randomly assigned to one of the four tutoring conditions. Participants returned after a two-day delay to complete a posttest CE assessment.

Participants in each condition were exposed to information using the same method of presentation via AutoTutor. One animated agent described a study. The other agent then instructed the participant to type in their answer to the following question: "Please describe and explain problems with the experiment, findings, or interpretation. If there are no problems, please type 'good experiment'." Although the instructions suggested that some studies might not contain any flaws, each study contained three or more flaws. The remaining interchange depended on condition. In the *full dialogue* condition, AutoTutor's curriculum script provided pumps, hints and prompts to guide the tutorial exchange beginning with the participant's initial answer. After all of the expectations had been covered, the tutor provided a summary of all

flaws for that particular problem in the form of a summary paragraph. In the *limited dialogue* condition, participants were only given the initial opportunity to correctly answer the question and, based on their answer, they were given immediate verification feedback. Participants were not given pumps, hints or prompts. Then a summary of the correct expectations was presented. In the *mixed dialogue* condition, the first three problems were presented with a "full dialogue" interaction and the last three problems were presented with the "limited dialogue" interaction. In the *no- dialogue* condition, participants simply watched and listened as AutoTutor's animated agents read the problem and the summary of correct expectations. Participants were not asked to provide an answer.

For all conditions, participants used headphones to listen to AutoTutor's animated agents. The text of the conversation was present on the screen at all times for all participants, allowing participants to read along with the agent voices and to review the material. All tutoring sessions were presented using personal computers in a private, sound-attenuated room. AutoTutor recorded the amount of time that students interacted with the tutor.

2.2. Results

Learning in each tutoring condition was assessed by performance on the CE pre and post assessment described above. To be scored as a correct identification of a flaw, participants could have either named the flaw or described it. For example, in an experiment description that did not contain a control group (i.e., a flaw), participants would be credited for identifying the flaw by simply stating that there was no control group, or acceptable synonyms (e.g., comparison group). Credit would also be given for describing the flaw (e.g., "there was no group that did not get any treatment"). Two raters who were blind to condition but trained in the procedure scored

the assessments (Agreement = 93%, Cohen's Kappa = .82). Any disagreements were discussed to the point of being resolved.

Performance scores were calculated for each participant and assessment (pre and post) by dividing the number of correctly identified flaws by the number of possible flaws. Table 2 shows the mean performance and efficiency scores as well as the average time spent during the tutoring session. Any post hoc tests were conducted with Tukey's test and an alpha level of .05.

2.2.1. Pretest performance. Pretest scores were analyzed to determine whether the groups were equally skilled prior to the tutorial. A one-way ANOVA on pretest scores with condition as a between-subjects variable produced no significant differences, $F(3,128) = 1.453$, $p > .20$, $MSE = .038$, $\eta^2 = .02$. To address any concerns regarding whether or not the two counterbalanced forms were equivalent assessments, pretest scores were entered into an independent samples t-test with form (A or B) as a grouping variable. This contrast was non-significant indicating that the tests were statistically equivalent, $t(127) = 1.54$, $p > .1$, $d = .27$

2.2.2. Overall impact of training on performance. A 2 (time-of-test: pre or post) X 4 (condition) mixed ANOVA with condition as a between subject variable produced a significant effect of time-of-test, indicating that overall, participants did learn from the tutoring session, $F(1,125) = 132.2$, $p < .05$, $MSE = .023$, $\eta^2 = .29$. Pairwise *t*-tests performed on pre versus posttest scores within each condition all produced significant results (all *p*'s < .05).

2.2.3. Posttest adjusted for pretest scores. An ANCOVA on posttest scores with pretest scores entered as a covariate and condition as the between-subjects factor assessed differences among conditions. The results indicated that there was a significant main effect of condition, $F(3,124) = 2.649$, $p = .05$, $MSE = .072$, $\eta^2 = .05$. As shown in Table 2, only the mixed dialogue condition was significantly different from the no-dialogue control condition. We tested

the Dialogue Hypothesis (Hypothesis 1) with a Jonckheere-Terpstra test of monotonic trends for performance based on difference scores (pretest subtracted from posttest). The order for testing Hypothesis 1 was full dialogue > mixed dialogue > limited dialogue > no-dialogue. The Jonckheere-Terpstra test was significant, $J = 2335$, $p < .05$ (one tailed), $z = -3.29$, $\eta^2 = .08$. The negative *z*-score indicates that as the amount of dialogue decreased across conditions, the learning gains of those conditions decreased. From an inspection of the means, however, it is evident that the pattern described by the Dialogue Hypothesis (Hypothesis 1) did not emerge because the full dialogue condition did not result in the highest performance.

2.2.4. Time-on-task. The time each participant interacted with the tutor was recorded and entered into a one-way ANOVA with condition as the between-subjects factor. The analysis produced a significant effect, $F(3, 125) = 223.05$ $p < .05$, $MSE = 8732.3$, $\eta^2 = .84$. The assumption of homogeneous variance among groups was violated as indicated by the significance of the Levene statistic ($p < .05$) due to the fact that there was little variance in the no-dialogue condition. However, considering that the groups were approximately equal (i.e., largest/smallest < 1.5, Stevens, 2007), the ANOVA results are robust to this violation. The results shown in Table 2 revealed that all conditions significantly differed from one another and conditions containing more dialogue took longer than those with less dialogue.

2.2.5. Efficiency score. Efficiency scores were computed by dividing the difference score (subtracting pretest from posttest percentage scores) by time-on-task (in minutes). For example, if a participant scored 25% on the pretest and 55% on the posttest and they interacted with the tutorial for fifty minutes, the equation would be (55-25)/50 and would produce an efficiency score of .6. This coefficient is the rate of change of performance to effort. Higher scores will indicate more efficient learning. The efficiency scores were entered into a one-way

ANOVA with condition as the between-subjects variable. The analysis produced a non-significant effect, $F(3, 125) = .732$ $p > .05$, $MSE = .281$, $\eta^2 = .02$.

One interpretation of this non-significant effect is that it provides no support whatsoever for the Diminishing Returns Hypothesis (Hypothesis 2). However, we should note that these conditions varied wildly on the amount of dialogue moves ranging from none to a full dialogue. As mentioned earlier, another approach to take is a more specific test of the hypothesis by comparing the conditions that had the same dialogue management. We tested the Diminishing Return Hypothesis by testing the difference between the mixed dialogue and full dialogue conditions by using a planned comparison. These two conditions used the same dialogue management structure but differed on the number of problems containing dialogue (mixed: 3 of 6 problems, full: 6 of 6 problems). The results were marginally significant, $t(125) = 1.95$, $p = .056$ (one-tailed), $d = .49$. As shown in Table 2, the mixed dialogue condition produced the most efficient results, providing support for the Diminishing Returns Hypothesis (Hypothesis 2).

2.3. Discussion

The results of Experiment 1 show that a natural language dialogue ITS can be made more efficient, but the picture is not simple. The Dialogue Interactivity Hypothesis was partially supported in that conditions containing dialogue resulted in more learning than conditions with little or no dialogue. The monotonic trend analysis suggested that increased dialogue among conditions increased performance. A closer inspection of the means in Table 2, however, revealed that the full dialogue condition failed to perform significantly different from any of the other conditions. The only significant difference was that the mixed condition outperformed the control condition.

In contrast, the results support the Diminishing Returns Hypothesis which states that the learning gains as a result of a full dialogue will decrease in size with increased time resulting in a decrease in efficiency. As predicted, engaging in only 3 full dialogue problems followed by 3 limited dialogue problems (mixed dialogue condition) was more efficient than the engaging in 6 full dialogue problems (full dialogue condition).

One observation of the data from this experiment was the fact that the no-dialogue condition produced an efficiency score equivalent to that of the mixed dialogue condition (see Table 2). Participants learned less in the no-dialogue condition than participants in the mixed condition, but they did so in less time, leading to equivalent efficiency scores. One must keep in mind, however, that the time in the no-dialogue condition was fixed across all participants. Participants in that condition watched the agents deliver a fixed script. If the time-on-task was extended in this condition to be similar to other conditions, and learning gains for this condition remained the same, the efficiency of this condition would decrease. Thus, Experiment 1 points to the need to equate the time on task to have a more valid and meaningful test of the Diminishing Returns Hypothesis.

3. Experiment 2

The results from the first experiment suggest that a condition containing dialogue could be made to be more efficient, but what about the learning due to time-on task? Was the difference in learning gains between the no-dialogue and mixed dialogue conditions simply due to less time on task in the former condition? According to the Dialogue Interaction Hypothesis (Hypothesis 1), increasing the time of exposure to the materials would not lead to improved learning gains if there was not an increase in quality dialogue. To explore this idea, a revised no-

dialogue condition was tested. In addition to their time spent with a no-dialogue tutoring session, participants were required to participate in an extra task inspired by the paradigm used to test for the generation effect (DeWinstanley & Bjork, 2004). Students in this condition were asked to copy identified expectations from the summary. This task is very similar to the type of note taking that is common in school. This procedure ensures that students are attending to the key concepts in the absence of self-generation.

Modifying the control condition in Experiment 2 limits comparisons across experiments but we can assess the pattern of data within experiments. We believe that this new control condition represents a very conservative control for testing the Dialogue Interactivity Hypothesis. It forces students to reread the summary information to the problems and increases time-on-task. This new control condition will be referred to as the 'no-dialogue-summary' condition.

In the second experiment, we also explored whether learning is affected by the ordering of the dialogue formats in the mixed dialogue condition (i.e., full versus limited). It is possible that the full dialogue in the first three problems acted as a model for question answering, and served to scaffold the process of evaluating the scientific studies (Collins, Brown, & Holum, 1991). Fading then occurred when the dialogue-rich problems were replaced by the limited dialogue interaction in which the student could apply what they learned during the first half of the session without engaging in a dialogue. On the other hand, beginning with the limited dialogue cases in which the student has to try to identify flaws without modeling presents a challenge that may help the student later learn better from the modeling (Schwartz & Bransford, 1998). By first attempting to identify the type of flaws on their own, students' may develop a more differentiated sense of the parameters of the scientific reasoning problems which may help

improve the effectiveness of later dialogues. The limited-full ordering exposure may act similar to the Schwartz and Bransford "Time for Telling" effect (1998). To test the generalizability of the Diminishing Returns Hypothesis (Hypothesis 2), we continued to use the mixed dialogue condition as in Experiment 1, but we added a new 'mixed dialogue (reversed)'condition in which limited dialogue was use for the first three problems and full dialogue for the last three.

The four tutorial conditions for the second experiment were: full dialogue, mixed dialogue, mixed dialogue (reversed), and no-dialogue-summary. According to the Dialogue Hypothesis (Hypothesis 1), performance will be dependent on the amount of dialogue interaction that occurs during the tutorial. The predicted pattern of performance will remain the same, but no differences are expected between the mixed dialogue conditions because they contain the same amount of dialogue (full > mixed = mixed (reverse) > no-dialogue-summary). According to the Diminishing Returns Hypothesis (Hypothesis 2) and the results of Experiment 1, both mixed conditions will produce the most efficient learning. To test this hypothesis, the efficiency scores of each of the mixed conditions were tested against the full dialogue condition in a planned contrast.

3.1. Method

3.1.1. Participants. The participants were 180 native English-speaking undergraduate introductory psychology students from a large Midwestern University who participated in this experiment for course credit. Ten participants did not return for the second experimental session. One participant was unable to complete the study due to computer failure. The attrition rates by condition were: no dialogue = 4, mixed dialogue = 2, mixed dialogue (reversed) = 2, full dialogue = 3. This left a total of 169 participants. None of the participants in the first

experiment were participants in the second experiment. As in Experiment 1, the knowledge level of these students about research methods was at the novice level (see Table 3).

3.1.2. Materials. The tutor content was the same as the four-flaw versions of the problems used in Experiment 1. The critique experiment assessments (CE) were also the same as Experiment 1.

3.1.3. Procedure. The procedure was exactly the same as in the first experiment (a two session study with pretest and tutor interaction on the first day followed by a two day delay and a posttest).

3.1.4. Conditions. The four conditions were: full dialogue, mixed dialogue, mixed dialogue (reversed), and no-dialogue-summary. The full dialogue and the mixed dialogue conditions were exactly the same as it was for the first experiment. The mixed dialogue (reversed) condition differed from mixed dialogue in that participants were first exposed to three problems presented in the "limited dialogue" followed by three problems presented in the "full dialogue" (L, L, L, F, F, F). Each concept was presented an equal number of times in a "full dialogue" method and in the "limited dialogue" method, as in Experiment 1. For the no-dialogue-summary condition, the initial interaction with the tutor was the same as in the first experiment. However, after the tutor was finished, each participant was presented with a paper copy of the summaries to the 6 problems that they just heard. The statements of the target expectations (i.e., flaws) in the summaries were underlined. Participants were required to rewrite the underlined expectation from these summaries onto a blank piece of paper provided to them by the experimenter. Thus, all groups wrote a similar amount of information, but in the no-dialogue-summary condition, participants did not generate the information through a student-tutor dialogue.

3.2. Results

Learning in each tutoring condition was assessed by performance on the CE task as in the first experiment. The percent agreement between the two raters was at 92% (Cohen's Kappa = .88). Any disagreements were discussed and resolved. Performance and efficiency scores and were calculated as they were for the initial experiment (pre and post). All means can be seen in Table 3. All post hoc tests were done with Tukey's test and an alpha level of .05.

3.2.1. Pretest performance. A one-way ANOVA on pretest scores with condition as a between-subjects variable produced no significant differences among the four conditions, $F(3,168) = .038$, $p > .500$, $MSE = .028$, $\eta^2 = .01$.

3.2.2. Overall performance. A 2 (time-of-test: pre or post) X 4 (condition) mixed ANOVA with condition as a between-subject variable produced a significant effect of time of test, indicating that overall, participants learned from the tutoring session, $F(1,165) = 53.5$, $p < .001$, $MSE = .022$, $\eta^2 = .43$. Pairwise t-tests performed on pre and posttest scores within each condition all produced significant results (all p's $< .05$).

3.2.3. Posttest adjusted for pretest scores. An ANCOVA conducted on posttest scores with condition as a between-subjects variable and pretest scores entered as a covariate indicated a main effect of condition, $F(3,168) = 3.736$, $p < .05$, $MSE = .039$, $\eta^2 = .06$. As shown in Table 3, all three experimental conditions had significantly higher learning gains than the no-dialogue-summarize control condition. A test of monotonic trends was conducted to test the Dialogue Hypothesis (Hypothesis 1) with the conditions in the following order: full dialogue > mixed dialogue = mixed dialogue (reverse) > no-dialogue-summary. The Jonckheere-Terpstra test was significant, $J = 3797$, $p < .05$ (one tailed), $z = -4.38$, $\eta^2 = .11$. The negative z-score indicates that as the amount of dialogue decreased across conditions, the learning gains of those conditions

decreased. However, looking at the means of conditions in Table 3, it is clear that the full dialogue condition did not produce the largest gains. All three experimental conditions had significant learning gains over the no-dialogue-summarize control condition, but none of the conditions containing dialogue significantly differed.

3.2.4 Time-on-task. The time that each participant spent during the tutoring sessions was recorded and entered into a one-way ANOVA. The results indicated a significant effect, $F(3, 165) = 26.29$ $p < .05$, $MSE = 2036.65$, $\eta^2 = .32$. The results shown in Table 3 revealed that the full dialogue condition took significantly longer than all other groups, which did not significantly differ. The added copying task was effective in increasing the time that the no-dialogue-summary participants spent on the target material.

3.2.5. Efficiency scores. The efficiency scores were calculated as they were for Experiment 1 and entered into a one-way ANOVA with condition as the between-subjects variable. The analysis produced a significant effect, $F(3, 165) = 4.116$ $p < .05$, $MSE = .561$, $\eta^2 = .07$. The Diminishing Returns Hypothesis (Hypothesis 2) was tested by comparing the efficiency scores of each of the mixed dialogue conditions to the full dialogue condition because these conditions contained the same dialogue management. The analysis of the mixed and the full dialogue conditions produced significant differences, $t(165) = 2.04$, $p < .05$ (one-tailed), $d = .46$; however the analysis of the mixed (reversed) and the full dialogue failed to produce significant differences between the means, $t(165) = .92$, $p > .05$ (one-tailed), $d = .22$. As shown in Table 3, it is clear that the mixed dialogue condition produced the most efficient results, providing support for the Diminishing Returns Hypothesis (Hypothesis 2). Although the mixed dialogue (reverse) failed to produce a significant difference from the full dialogue, the means of the conditions are consistent with the prediction of Hypothesis 2.

3.3. Discussion

The pattern of results from Experiment 2 is fairly clear: the mixed dialogue conditions performed as well or better than a full dialogue condition and took less time. The Dialogue Interactivity Hypothesis (Hypothesis 1) was supported because conditions containing a quality dialogue interaction outperformed one in which no dialogue took place. However, the support was only partial because the full dialogue condition failed to outperform all partial dialogue conditions. The hypothesis was also supported by the finding that increasing the time that the no-dialogue participants spent on the material did not increase learning to the level of the dialogue conditions. Thus, it is not just time but the quality of instruction during that time that matters most.

The efficiency scores indicated that the full dialogue condition produced the least efficient scores when compared to other conditions containing a quality dialogue interaction. This supports the notion that time spent in a dialogue produces a diminished return with regard to learning (Hypothesis 2). Thus, reducing the number of problems containing a full dialogue interaction during a tutoring session could lead to more efficient learning.

The results also suggest the ordering of the dialogue formats in the mixed dialogue problems had a modest though non-significant influence on learning. Both mixed conditions led to greater learning gains than the no-dialogue summary condition but they did not differ from each other. These findings should be encouraging for designers of learning environments and ITSs which rely on dialogue between computer and user. The fact that there were no differences in this experiment among these conditions suggests that designers may have some freedom in choosing a dialogue format (full or limited) for a given problem without sacrificing learning. The designer or teacher can determine the order based on other constraints or even allow the

student to select the order, which may increase motivation (Malone & Lepper, 1987). However, it may be beneficial to explore this issue further in a future study.

4. General Discussion

The goal of this study was to investigate the efficient use of natural language dialogue during a computerized tutoring session. This is the first study of a natural language based ITS that systematically manipulated conditions to assess learning efficiency, as opposed to merely measuring efficiency among existing treatments (i.e., Author/s, 2007[a]). The central question is not whether such systems produce significant learning gains above less interactive conditions because it is already well documented that this occurs (Author/s, 2001; Author/s, 2004[a]; Author/s, 2007[a]; Author/s, 2007[b]). Rather it is a question of whether or not such systems could be made to be more efficient. The present results suggest that there is a point during a tutoring session when the benefit of engaging in a tutorial dialogue results in a diminished return of the investment of time.

The results do not support a strong form of the Dialogue Interactivity Hypothesis (Hypothesis 1). Incrementally increasing the amount of dialogue did not incrementally lead to more learning of that topic. These results do, however, support a weaker form of the hypothesis which is that some dialogue is better than no dialogue at all. In Experiment 1, only the mixed dialogue condition outperformed the no-dialogue condition and in Experiment 2 all dialogue conditions led to better learning than the no-dialogue-summary condition.

When looking specifically at conditions containing a quality interaction between the tutor and the student, a more streamlined condition containing both a full and limited dialogue interaction was the most efficient, providing support for the Diminishing Returns Hypothesis

(Hypothesis 2). The results of these experiments suggest that there may be an optimal amount of dialogue for particular learning tasks.

There are several potential reasons for the diminished returns from these full dialogue interactions. First, it may be that the student has mastered the material. If this had occurred, then having extra time would not be needed and would only decrease the efficiency scores. Obviously, this was not the case for this study considering the posttest performance scores were not at ceiling for any condition. However, it may be that these students learned as much as they could during their one exposure to the tutor and that students would benefit in subsequent exposures to the full dialogue AutoTutor. A second possibility is that the participants needed additional information to understand the targeted concepts. If this occurred, additional dialogue would not be beneficial unless it contained the information the student needed to fully grasp the concepts. A third reason for diminished returns in the full dialogue condition would be fatigue. The students in the full dialogue condition may have simply gotten bored or tired of engaging in a large number of problems with dialogues. This could have led to a loss of motivation and disengagement from the task, leading to little new learning once this point had been reached.

Prior research does suggest that the fading of support may be more beneficial for learning than continuing to give complete support. A recent study by Nückles, Hübner, Dümer, and Renkl (2010) found that fading metacognitive prompts in a journal writing activity to be more beneficial than continuous prompting. In their study, students who received continuous prompting over a period of time performed more poorly than students for whom the prompting process was faded. These results suggest that receiving continual instructional support may interfere with the learning process in that it does not allow for students to apply the strategies they learned. Furthermore, it may be that the prompting has a negative impact on motivation. In

the same study, Nückles et al. also found that continuous prompting decreased motivation over time according to self-report motivational measures.

The conditions of the current study are similar to the Nückles et al. (2010) study in that the mixed dialogue condition faded the prompting while participants in the full dialogue condition continuously engaged in a dialogue that prompted them to produce more information. This continual prompting could have also impeded the learning process because it may not have allowed students to apply the strategies that they learned to new problems. We did not include a measure of interest or engagement in the current study, so it is difficult to gauge the motivation levels of participants in different conditions. Nevertheless, exploring ways to minimize the unproductive use of cognitive resources in ITS environments seems potentially beneficial when developing and testing such systems.

Finally, it may be the case that the diminished effectiveness of additional dialogues was due to a limitation of the tutoring system itself. AutoTutor tries to match the student input with pre-scripted expected answers using word-matching algorithms and LSA, and it is this match that partially determines the next dialogue move that AutoTutor performs. While the word-matching algorithms and LSA are very effective, no natural language-based system is perfect. So, it may be that the student provides a conceptually correct answer, but in some cases AutoTutor may fail to accept it as a match. When this happens AutoTutor may employ a pump, hint, or prompt. To the extent that this occurs, these unnecessary moves add to the length of the tutoring session and probably do not add anything in terms of learning.

While we expect that the point of diminished returns could be identified for most ITS's, we believe it is a complex issue that depends on many different factors. The point in which people reach diminishing returns will likely be based on the complexity of the material, the

knowledge of the student, their willingness to learn, the number of learning sessions, and the quality of the dialogue. It is not possible to address all of these factors in one study. In the current study, we had novices learning a mixture of deep and shallow knowledge in a single session. We might find support for the stronger version of the Dialogue Interactivity Hypothesis if we replicate this study with more intermediate students or novices learning across multiple sessions. Additionally, the CE assessment could be considered near knowledge transfer as it was very similar to the problems experienced during the tutoring session. A far transfer task may better illustrate the beneficial effect of a full dialogue tutoring session because in a far transfer task, a student would have to engage in a deeper level of reasoning. AutoTutor enables a deeper level of reasoning because the dialogue promotes deeper levels of knowledge construction.

There are several possibilities regarding future studies based on these results. One interesting idea may be to examine different orderings of mixed and full dialogue such as interspersing. Although the ordering of mixed and full problems did not appear to matter in the current experiment, future experiments could more carefully test learning gains at the transition point to determine when learning occurred. One could also modify an ITS to determine the point for fading based on individual performance within a session. A second interesting possibility would be to test different control conditions than the ones we employed here. In the current study, we tried to keep everything constant across conditions while varying the amount of dialogue. As such, our control conditions required that the students listen as the animated agents read the problem and summary statements to them. However, it may be possible that a control condition in which the students read the information may create an environment in which the students are more engaged and possibly take a more active role in the process of knowledge acquisition.

Overall, these results should be encouraging for educators and students. Educators could use such programs and cover more material in a learning session. We have shown that by mixing full dialogue problems with limited dialogue ones, students can be exposed to more example problems without sacrificing learning. The fact that boredom negatively correlates with learning implies that students would benefit from an environment that reduces the risk of them losing interest (Author/s, 2004[b]; Author/s, 2009[b]). The longer a student spends in a learning environment, the greater the risk of the student getting bored, unmotivated, unengaged and frustrated. Therefore, any factor which would reduce time-on-task while preserving learning gains should be rigorously pursued.

With advancements in natural language understanding and the increased availability of ITSs, there has been an increase in research regarding the components within these systems and their functionality (Author/s, 2007[c]; Deiglmayer & Spada, 2011; Lane & VanLehn, 2005; Makitalo-Siegl, Kohnle, & Fischer, 2011; Moreno, 2009; Roll, Aleven, McClaren, & Koedinger, 2011). Given this increase, it is even more critical that researchers and designers examine issues of time and efficiency of learning. It is important to identify efficient ways to incorporate such systems so that they will continue to be beneficial while holding costs (time, money for development) to a minimum. The current study shows one way in which intelligent tutoring systems containing natural dialogue can be made to be more efficient. We believe future research will identify other ways and this endeavor will ultimately make ITS's more practical for use in the classroom.

Acknowledgements

The research on was supported by the National Science Foundation (ITR 0325428, ALT-0834847, DRK-12-0918409) and the Institute of Education Sciences (R305B070349, R305A080589, R305A080594). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF or IES. These experiments were conducted in partial fulfillment of a Master's degree for first author. Thanks to committee members Joseph Magliano and Patricia Wallace for their feedback. Thanks also to Katja Wiemer for her help preparing experimental materials. Lastly, thanks to Patrick Chipman for his help with the computer programming.

Appendix

An actual student-tutor dialogue for a scientific reasoning problem to illustrate the different components of AutoTutor. The bolded information classifies important tutor responses.

Tutor: I am not ever going to jog because I read in a health magazine that running increases stress levels. It described an experiment that showed just that. In this experiment, participants were placed into either a low stress group or a high stress level group, based on their pre-existing stress levels. There were five participants in each group. Both groups were instructed to run three times a week. The result was that the stress levels in the high stress level group increased even more. They were really stressed. The stress level in the low stress group stayed about the same. So you see, doing exercise leads to higher stress, at least for people with some stress in their lives. However, I think there are some problems with the experiment and its description.
<Problem>
Please describe and explain to me problems with the experiment, findings, or interpretation. If there are no problems, type good experiment.
Student: There was no control group.
Tutor: In other words, you've pointed out that
Tutor: This study contained no control group since all the participants jogged. **<Assertion>**
Tutor: I see, ok, umm,
Tutor: And can you add to that? <**Pump>**
Student: You need to know how to define stress in order to measure it.
Tutor: Perfect answer! <**Positive Feedback>**
Tutor: Stress was not operationally defined. <**Assertion>**
Tutor: Let's try this together.
Tutor: What can you say about the number of people included in this sample? <**Hint>**
Student: There were too few.
Tutor: Okay. <**Neutral Feedback>**
Tutor: Now
Tutor: What could be problematic with the sample size? <**Hint>**
Student: It may not be representative of the population. You may not be able to generalize from such a small sample size.
Tutor: I think you've covered something important here.
Tutor: The size of the sample is too small to generalize the results. <**Assertion>**
Tutor: If some factors were not accounted for, then tell me what might be a problem with this experiment? **<Hint>**
Student: There may be confounding variables.
Tutor: Yeah! **<Positive Feedback>**
Tutor: I bet you can get this!
Tutor: Other factors that may affect the response are called confounding what? **<Prompt>**
Student: Variables.
Tutor: Good answer! <**Positive Feedback>**
Tutor: You've got that. However, there's something you haven't mentioned.
Tutor: You know
Tutor: What can you tell us about the differences between the high and low stress groups?

<**Hint**>
Student: They were either depressed or not depressed before the experiment even began.
Tutor: Yeah... <**Neutral Feedback**>
Tutor: Recall that
Tutor: Confounding variables may be causing the change in stress level. <**Assertion**>
Tutor: Look at it this way:
Tutor: Considering how the independent variable was carried out, what type of study or research design is used here? <**Prompt**>
Student: Correlational study?
Tutor: Your question is good, but what else can you say? <**Pump**>
Tutor: I'm sorry, let me repeat,
Tutor: Considering how the independent variable was carried out, what type of study or research design is used here? <**Prompt**>
Student: Quasi-experiment.
Tutor: Yeah, that's right! <**Positive Feedback**>
Tutor: Let's try this.
Tutor: What type of study merely measures the association between two or more variables? <**Prompt**>
Student: Correlational study.
Tutor: Good. <**Positive Feedback**>
Tutor: We've almost covered it, but there's one part I think is still missing.
Tutor: Consider this.
Tutor: Think about what is required for a true experiment. What does that tell you about the design? <**Hint**>
Student: Internal validity is required for a true experiment.
Tutor: possibly, <**Neutral Feedback**>
Tutor: It is significant that
Tutor: This is a correlational study and not an experiment. <**Assertion**>
Tutor: Now, let's recap.
Tutor: There could be confounding variables that are responsible for the difference on stress levels between the high and low stress groups. Confounding variables are other variables besides the independent variable that systematically affects the results of a study. For example, jogging is an activity that takes time and which could increase stress in individuals who are already stressed. So, it might not be the jogging that increased the stress in the high stress group but the fact that there was another task to perform. Also, technically this was not a true experiment as claimed by Joe because the groups were not formed by the experimenter but were based on their pre-existing stress levels. The sample size also appears to be too small for meaningful results. Perhaps most importantly, there was no control group in which individuals did not jog, and this would be needed to support the conclusion that jogging causes stress. Lastly, there was no operational definition of stress, and therefore, it is hard to know exactly what was being measured. Operational definitions provide a clear and exact way of measuring the concept of interest, and in this case, it was stress. **<Summary> <Explanatory/Corrective Feedback>**

References

Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science, 26*, 147-179. doi: 10.1016/S0364-0213(02)00061-7

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*(2), 167-207. doi: 10.1207/s15327809jls0402_2

Atkinson, R. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology , 94*, 416 - 427. doi:10.1037//0022-0663.94.2.416

Author/s (1994).

Author/s (1995).

Author/s (2001).

Author/s (2003).

Author/s (2004[a]).

Author/s (2004[b]).

Author/s (2005).

Author/s (2007[a]).

Author/s (2007[b]).

Author/s (2007[c]).

Author/s (2008).

Author/s (2009[a]).

Author/s (2009[b]).

Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents *International Journal of Artificial Intelligence in Education, 15*, 5–115.

Carroll, J. (1963). A model of school learning. *Teachers College Record*, *64*, 723-733.

Carroll, J. B., & Spearritt, D. (1967). A study of a "model of school learning." Monograph, Number 4. Center for Research and Development on Educational Differences, Harvard University, 1967.

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477. doi:10.1207/s15516709cog1803_3

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471-533. doi:10.1016/S0364-0213(01)00044-1

Cobb, C. W., and Douglas, P. H. (1928). A Theory of Production. *American Economic Review Supplement, 23*, 139-65.

Cole, R., van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., & Movellan, J., (2003). Perceptive animated interfaces: First steps toward a new paradigm for human computer interaction. *Proceedings of the IEEE*, *91*, 1391-1405. doi:10.1109/JPROC.2003.817143

Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator, 6-11*, 38-46.

Deiglmayer, A., & Spada, H.(2011). Training for fostering knowledge co-construction from collaborative inference-drawing. *Learning and Instruction, 21*(3), 441 – 451. doi:10.1016/j.learninstruc.2010.06.004

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generations effect: Implications for a better reader. *Memory and Cognition, 32* (6), 945-955. doi:10.3758/BF03196872

Fox, B. (1993). *The human tutorial dialogue project.* Hillsdale, NJ: Erlbaum.

Frederick, W., & Walberg, H. (1980). Learning as a Function of Time. *Journal of Educational Research, 73 (4)*, 183-94.

Hoffman, B., & Schraw, G. (2009). The influence of self-efficacy and working memory capacity on problem-solving efficiency. *Learning and Individual Differences, 19*, 91–100. doi:10.1016/j.lindif.2008.08.001

Hoffman, B., & Shraw,. G. (2010). Conceptions of Efficiency: Applications in Learning and Problem Solving. *Educational Psychologist, 45,* 1-14. doi:10.1080/00461520903213618

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education, 11*, 47-78.

Karweit, N., & Slavin, R. E. (1981). Measurement and Modeling Choices in Studies of Time and Learning. *American Educational Research Journal, 8,* 157 – 171. doi:10.2307/1162379

Karweit, N. L. (1984) Time-on-Task Reconsidered: A Synthesis of Research on Time and Learning. *Educational Leadership, 41*, 33-35.

Lane, H. C., &VanLehn, K. (2005). Teaching the tacit knowledge of programming to novices with natural language tutoring. *Computer Science Education, 15*(3), 183–201.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284. doi:10.1080/01638539809545028

Litman, D.J, Rose, C.P., Forbes-Riley, K., VanLehn, K., Bhembe, D., and Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, *16*, 145-170.

Makitalo-Siegl, K., Kohnle, C., & Fischer, F. (2011). Computer-supported collaborative inquiry learning and classroom scripts: Effects on help-seeking processes and learning outcomes. *Learning and Instruction, 21*(2), 257 – 266. doi:10.1016/j.learninstruc.2010.07.001

Malone, T. W., & Lepper, M. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning and instruction: Vol. 3. Conative and affective process analysis.* Hillsdale, NJ: Erlbaum.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers, 36*, 222 - 233. doi:10.3758/BF03195567

Moore, J. D. (1995). *Participating in explanatory dialogues*. Cambridge: MIT Press.

Moreno, R. (2009). Constructing knowledge with an agent-based instructional program: A comparison of cooperative and individual meaning making. *Learning and Instruction, 19*(5), 433 – 444. doi:10.1016/j.learninstruc.2009.02.018

Moreno, R., & Mayer, R. E. (2007). Interactive multimodal learning environments. *Educational Psychology Review, 19*, 309-326. Doi: 10.1007/s10648-007-9047-2

Nückles, M., Hübner, S., Dümer, S., & Renkl, A. (2010). Expertise reversal effects in writing-to-learn. *Instructional Science, 38*(3). 237 – 258. doi:10.1007/s11251-009-9106-9

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23,* 90-108. doi:10.1006/ceps.1997.0959

Roll, I., Aleven, V., McClaren,B., & Koedinger, K. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267 – 280. doi:10.1016/j.learninstruc.2010.07.004

Rossi, P. H., Lipsey, M.W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Schwartz, D., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*, 475-522. doi:10.1207/s1532690xci1604_4

Shah, F., Evens, M., Michael, J., & Rovick, A. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes, 33*, 23-52. doi:10.1207/S15326950DP3301_02

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational research, 78*, 153 – 189. doi 10.3102/0034654307313795

Slamecka, N.J. & Graf, P. (l978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592-604.

Stevens, J. P. (2007). *A Modern Approach to Intermediate Statistics*. New York: Erlbaum.

Taraban, R., Rynearson, K., & Stalcup, K. A. (2001). Time as a variable in learning on the World-Wide Web. *Behavior Research Methods, Instruments, & Computers, 33*(2), 217-225. doi:10.3758/BF03195368

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. The *Journal of the Learning Sciences, 2*, 1-59. doi:10.1207/s15327809jls0201_1

VanLehn, K., Lynch, C., Taylor, L., Weinstein, A., Shelby, R., Schulze, K., Treacy, D., & Wintersgill, M. (2002). Minimally invasive tutoring of complex physics problem solving.

In S. A. Cerri, G. Gouarderes, & F. Paraguacu, (Eds.), *Intelligent Tutoring Systems: 6th International Conference* (pp. 367-376). Berlin: Springer.

Walberg, H. J. (1983). Science literacy and economic productivity in international perspectivity. Da*edalus,112,*1-28.

Walberg, H. J., & Tsai, S. (1984). Reading achievement and diminishing returns to time. *Journal of Educational Psychology, 76*, 442-451. doi:10.1037//0022-0663.76.3.442

Table 1

Example script of problem presenting a description of a study containing flaws relevant to the target experimental design concepts

Problem Statement
Agent 1 (Joe): I do not think I will buy any more textbooks. I read about this experiment done at a top University where they showed that students learn just as much regardless of whether or not they read the textbook. The experiment was conducted in two sections of a social psychology course. In the fall semester, all students in a specific course were told that the textbook was optional. In the spring semester, all students in the same course were told that reading the textbook was required. Both classes had ten students in each class. The same professor taught the two courses and gave the same lectures. The final exam for both classes was an essay test that was made from the lecture materials. After the completion of both classes, the professor compared the essay grades in the two classes. Just as the professor suspected, there was no statistical difference on the final exam scores between the two classes. The average grade was 94% for the fall class and 95% for the spring class. In fact, all students in both classes received better than 90% on the final exam. So, I think the textbook does not matter. And if it does not matter, why buy textbooks?

Agent 2 (Crystal): I am not so sure about that study. So I plan to continue reading my textbooks. Please describe and explain to me problems with the experiment, findings or interpretation. If there are no problems, type good experiment.

Expectation1: The sample size is too small to generalize results.
Expectation2: Experimenter bias could have contaminated the results.
Expectation3: The final may not have been scored objectively.
Expectation4: The dependent variable may not have been sensitive enough.
Expectation5: The test is not a valid measure of what could be learned from the textbook.

Example hint for expectation1:
Hint: What could be problematic about the number of students sampled in this study?
Expected answer: The sample size is too small to make a generalizable conclusion.

Example prompt for expectation1:
Prompt: When you use a representative portion of the population, you do not want the number of participants to be what?
Expected answer: small

Assertion for expectation1:
Because the sample size was small, the results may not generalize to the population.

Table 2

Experiment 1: Performance means for pretest, posttest, adjusted post for pretest (M = .25), average time-on-task (minutes), and efficiency scores by condition with standard deviations in parentheses

	No Dialogue	Limited Dialogue	Mixed Dialogue	Full Dialogue
Number of participants	32	33	34	30
Pretest[1]	.21 (.17)[a]	.27 (.15)[a]	.24 (.16)[a]	.29 (.18)[a]
Posttest[1]	.31 (.21)[b]	.38 (.17)[a, b]	.45 (.20)[a]	.41 (.22)[a, b]
Posttest[1] (adjusted for pretest)	.33 (.19)[b]	.37 (.19)[a, b]	.46 (.22)[a]	.39 (.19)[a, b]
Time-on-task[1] (minutes)	23 (.42)[d]	33 (5.48)[c]	50 (8.01)[b]	60 (7.91)[a]
Efficiency Scores[2]	.42 (.93)	.34 (.58)	.42 (.40)[a]	.21 (.42)[b]

Notes:

[1] Common superscript denotes non-significant differences assessed by Tukey's test and an alpha level of .05.

[2] Common superscript of efficiency scores denotes non-significant difference among planned comparisons between the mixed dialogue condition and the full dialogue condition.

Table 3

Experiment 2: Performance means for pretest, posttest, adjusted post for pretest (M = .23), average time-on-task (minutes), and efficiency scores by condition with standard deviations in parentheses

	No Dialogue-summarize	Mixed Dialogue (reverse)	Mixed Dialogue	Full Dialogue
Number of participants	40	43	44	42
Pretest[1]	.24 (.18)[a]	.23 (.17)[a]	.23 (.18)[a]	.23 (.14)[a]
Posttest[1]	.34 (.22)[b]	.43 (.20)[a, b]	.48 (.24)[a]	.43 (.19)[a, b]
Posttest[1] (adjusted for pretest)	.34 (.20)[b]	.43 (.20)[a]	.48 (.20)[a]	.43 (.19)[a]
Time-on-task[1] (minutes)	59 (8.24)[a]	58 (8.24)[a]	56 (7.22)[a]	71 (11.09)[b]
Efficiency Scores[2]	.17 (.37)	.35 (.38)[b]	.44 (.42)[a]	.28 (.29)[b]

Notes:

[1] Common superscript denotes non-significant differences assessed by Tukey's test and an alpha level of .05.

[2] Common superscript denotes non-significant difference among planned comparisons between each mixed dialogue condition and the full dialogue condition.

Figure 1. A screenshot of AutoTutor with two animated pedagogical agents: Crystal (left) and Joe (right).