Student Prior Knowledge and Learning in an Intelligent Tutoring System: Comparing the Effectiveness of Vicarious and Interactive Dialogues

Keith Thomas Shubeck

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation

This Dissertation is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.
STUDENT PRIOR KNOWLEDGE AND LEARNING IN AN INTELLIGENT TUTORING SYSTEM: COMPARING THE EFFECTIVENESS OF VICARIOUS AND INTERACTIVE DIALOGUES

by

Keith T. Shubeck

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Major: Experimental Psychology

The University of Memphis
Acknowledgements

I would like to acknowledge my advisor, Xiangen Hu. He was a consistent source of inspiration to me throughout my graduate school experience. I especially appreciate his patience, wisdom, support, and friendship. He helped me grow as a person and as a researcher. I would like to thank Art Graesser, who also provided, and continues to provide, excellent advice and encouragement. I am very grateful for my other committee members, Phil Pavlik and Scotty Craig, both of whom played essential roles during my time as a graduate student.

I would also like to thank both my lab colleagues and the professors in the psychology department and the IIS, who always found the time to help me, without hesitation.

Finally, I would like to acknowledge my family. Their unconditional love and support provided the light I needed when in the dark and the foundation I needed to succeed.
Abstract

Shubeck, Keith. [Ph.D.]. The University of Memphis. [May 2022]. Student Prior Knowledge and Learning in an Intelligent Tutoring System: Comparing the Effectiveness of Vicarious and Interactive Dialogues. Major Professor: Xiangen Hu, PhD.

Intelligent tutoring system (ITS) developers often assume that the system’s effectiveness is driven by its interactivity, and for good reason; research regularly shows that interactivity is a critical component for learning. However, some evidence suggests that students with low prior knowledge learn best in vicarious settings, or in settings where they can model a peer agent’s positive learning behaviors. The expertise reversal effect suggests that high prior knowledge students may not benefit from the increased granularity and interactivity that a conversation-based ITS can provide.

In the current study, participants were randomly placed into an interactive condition where they interacted with AutoTutor, an ITS for critical thinking, or in a vicarious condition where they observed another user interact with AutoTutor.

Students who are observing a tutoring session can model positive learning behaviors of the tutee. However, the results indicated that low prior knowledge participants who watched high prior knowledge participants (LH) interact with AutoTutor was the worst performing subgroup. This suggests that observing more accurate and positive learning behaviors in AutoTutor was not enough to promote learning for the LH group. High prior knowledge participants who observed low prior knowledge participants (HL), and both the high and low prior knowledge interactive participants significantly outperformed the LH group on posttest scores. There was no support for the expertise reversal hypothesis. While the results were not significant, high prior knowledge participants who viewed other high prior knowledge participants (HH) had lower posttest scores on average than high prior knowledge participants who viewed low prior knowledge participants.
(HL). There was also no significant difference between the high prior knowledge interactive group and the low prior knowledge interactive group, which runs counter to what the expertise reversal hypothesis would predict. These findings suggest that the learning of high prior knowledge participants was not inhibited by the increased amount of feedback and potentially redundant information provided by low prior knowledge participants in the vicarious condition. Instead, the trend suggests that the increased amount of negative feedback, conflict episodes, and potential contradictions observed in vicarious settings may have benefitted high prior knowledge participants in vicarious settings.

**Keywords:** prior knowledge, intelligent tutoring system, conversation framework, vicarious learning
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>Chapter 1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 2. Human Tutoring: Theory and Research</td>
<td>5</td>
</tr>
<tr>
<td>Learning in Tutoring Theory</td>
<td>5</td>
</tr>
<tr>
<td>Human Tutoring Strategies &amp; Effectiveness</td>
<td>7</td>
</tr>
<tr>
<td>Question Asking in Tutoring</td>
<td>11</td>
</tr>
<tr>
<td>Self-Explanations</td>
<td>13</td>
</tr>
<tr>
<td>Vicarious Learning in Tutoring</td>
<td>16</td>
</tr>
<tr>
<td>The ICAP Framework</td>
<td>18</td>
</tr>
<tr>
<td>Chapter 3. Intelligent Tutoring Systems: Motivation, Design, and Effectiveness</td>
<td>24</td>
</tr>
<tr>
<td>Approaches</td>
<td>27</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>29</td>
</tr>
<tr>
<td>AutoTutor</td>
<td>32</td>
</tr>
<tr>
<td>Authoring AutoTutor Scripts</td>
<td>33</td>
</tr>
<tr>
<td>Tutoring Dialogue in AutoTutor</td>
<td>35</td>
</tr>
<tr>
<td>AutoTutor Effectiveness</td>
<td>38</td>
</tr>
<tr>
<td>AutoTutor and multimedia learning principles</td>
<td>39</td>
</tr>
<tr>
<td>Vicarious Learning in AutoTutor.</td>
<td>44</td>
</tr>
<tr>
<td>Trialogues in ITS</td>
<td>49</td>
</tr>
<tr>
<td>Operation ARIES</td>
<td>50</td>
</tr>
<tr>
<td>AutoTutor CSAL</td>
<td>51</td>
</tr>
<tr>
<td>Chapter 4. Research Questions &amp; Methodology</td>
<td>53</td>
</tr>
<tr>
<td>Research Questions</td>
<td>53</td>
</tr>
<tr>
<td>Conditions</td>
<td>54</td>
</tr>
<tr>
<td>Hypotheses</td>
<td>55</td>
</tr>
<tr>
<td>Materials, Participants, and Procedure</td>
<td>57</td>
</tr>
<tr>
<td>Content</td>
<td>57</td>
</tr>
<tr>
<td>Interface</td>
<td>58</td>
</tr>
<tr>
<td>Participants</td>
<td>59</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment Tools</td>
<td>60</td>
</tr>
<tr>
<td>Interaction Replays</td>
<td>62</td>
</tr>
<tr>
<td>Procedure</td>
<td>64</td>
</tr>
<tr>
<td>Chapter 5. Results</td>
<td>66</td>
</tr>
<tr>
<td>T- Sample Differences on Prior Knowledge and Assessment Times</td>
<td>66</td>
</tr>
<tr>
<td>MTurk &amp; SONA</td>
<td>66</td>
</tr>
<tr>
<td>Vicarious versus Interactive Conditions</td>
<td>68</td>
</tr>
<tr>
<td>Group Differences on Learning</td>
<td>69</td>
</tr>
<tr>
<td>Prior Knowledge on Post-Training, Posttest Scores</td>
<td>69</td>
</tr>
<tr>
<td>Effects of Knowledge on Learning Behaviors in AutoTutor</td>
<td>70</td>
</tr>
<tr>
<td>Condition Differences on Posttest Scores</td>
<td>72</td>
</tr>
<tr>
<td>Condition Differences with Prior Knowledge on Posttest Scores</td>
<td>73</td>
</tr>
<tr>
<td>Vicarious Pair Groups</td>
<td>76</td>
</tr>
<tr>
<td>Vicarious Domain Knowledge Distances Effect</td>
<td>79</td>
</tr>
<tr>
<td>Chapter 6. Discussion</td>
<td>83</td>
</tr>
<tr>
<td>Implications</td>
<td>90</td>
</tr>
<tr>
<td>Limitations</td>
<td>92</td>
</tr>
<tr>
<td>Future Directions</td>
<td>95</td>
</tr>
<tr>
<td>Conclusion</td>
<td>97</td>
</tr>
<tr>
<td>References</td>
<td>98</td>
</tr>
<tr>
<td>Appendix</td>
<td>117</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Hypotheses with Participant Domain Knowledge and References</td>
<td>55</td>
</tr>
<tr>
<td>2. Key Concepts and Examples of Definitional and Applied Questions</td>
<td>61</td>
</tr>
<tr>
<td>3. T-Tests on MTurk and SONA for Pretest, Post-Training, and Posttest</td>
<td>66</td>
</tr>
<tr>
<td>4. Means and Standard Deviations for Conditions on Pretest and Posttest Scores</td>
<td>72</td>
</tr>
<tr>
<td>5. Estimated Marginal Means of Posttest Scores as a Function of Pretest Subgroups with Post-Training as Covariate</td>
<td>75</td>
</tr>
<tr>
<td>6. Estimated Marginal Means of Posttest Scores as a Function of Post-Training Subgroups with Pretest as Covariate</td>
<td>75</td>
</tr>
<tr>
<td>7. Estimated Marginal Means of Posttest Scores as a Function of Pretest Subgroups with Pretest as Covariate</td>
<td>77</td>
</tr>
<tr>
<td>8. Estimated Marginal Means for Post-training Subgroups with Pretest as Covariate</td>
<td>78</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. AutoTutor Conversation Engine.</td>
<td>33</td>
</tr>
<tr>
<td>2. Four agents in the AutoTutor interface.</td>
<td>40</td>
</tr>
<tr>
<td>3. The AutoTutor interface.</td>
<td>59</td>
</tr>
<tr>
<td>4. An excerpt of an xAPI statement as recorded in the LRS.</td>
<td>62</td>
</tr>
<tr>
<td>5. A screenshot of an AutoTutor “replay” viewed by a vicarious participant.</td>
<td>63</td>
</tr>
<tr>
<td>6. Scatter Plot of Post-Training Scores on Number of Responses to AutoTutor.</td>
<td>71</td>
</tr>
<tr>
<td>7. Scatter Plot of Number of Responses on Posttest Scores.</td>
<td>72</td>
</tr>
<tr>
<td>8. Scatter Plot of Pretest Distance on Posttest Scores.</td>
<td>81</td>
</tr>
<tr>
<td>9. Scatter Plot of Post-Training Distance on Posttest Scores.</td>
<td>82</td>
</tr>
</tbody>
</table>
Chapter 1. Introduction

There is no “one size fits all” ideal learning environment. While this is especially true for environments without individualized interaction and adaptation (e.g., lecture-based classrooms) this also applies to certain conditions in one-on-one tutoring, which are inherently adaptive and individualized. The effectiveness of any learning environment will vary from student to student. Low-level students are frequently left behind in traditional lecture-based classrooms. For example, a significant portion of U.S. students score below basic proficiency levels for reading, mathematics, and science. According to the National Center for Educational Statistics, in 2017, 20% of 4th graders and 30% of 8th graders scored “below basic” for mathematics and 38% of 12th graders scored below basic levels in 2015 (McFarland et al., 2017). Alternatively, high-level students may be constrained by a curriculum that caters to medium-level students. It is well-established that active learning environments regularly outperform traditional lecture based classrooms (Burrowes, 2003; Freeman et al., 2014; J. Michael, 2006; P. Michael, 2004; Wilke & Straits, 2001). While the classroom environment may vary from school to school, they are primarily lecture-based (i.e., primarily vicarious, simple communication of knowledge) with an average student teacher ratio of 26.8:1 (Snyder, 2018). A solution to the one-size-fits all lecture approach is found in individual expert-led tutoring, which involve active learning and individualized learning. Tutoring is often considered the gold-standard of learning environments (Bloom, 1984; Graesser et al., 2001).

In tutoring, students are provided more opportunities to play a larger role in the construction of their knowledge (M. T. H. Chi et al., 2001; Lave, 1990). Comparatively, students in traditional classrooms have less opportunities for interaction with the learning material and are often simply passive receivers of information. For example, during tutoring students ask 26.7 questions per
hour compared to 0.11 individual learner questions in a classroom environment (Graesser & Person, 1994). Additionally, students who are tutored made an average of 206 statements during the span of about 2 hours (M. T. H. Chi et al., 2001). This level of individual interactivity is simply not possible in a classroom environment. Another advantage of tutoring is the ability for the tutor to adjust to the student’s needs. Tutors can focus on individual concepts that are troublesome for the learner and the student’s misconceptions, if detected, can be quickly addressed. However, individual tutoring is often expensive and is not available to students who may need it the most. Fortunately, the development of computer-based learning environments, and more specifically intelligent tutoring systems (ITSs), presents an opportunity to increase access to individualized and highly effective learning environments. Intelligent tutoring systems are a specific type of computer-based learning environment that intelligently adapt to individual students while applying pedagogical strategies that resemble human tutoring (Graesser, D’Mello, & Cade, 2011). However, the development of effective and truly adaptable ITSs presents a new set of challenges.

The adaptability of ITSs has improved over time but remains a challenge and a major focus for ITS research. There is a large history of research that explores aptitude treatment interactions (ATIs) in different learning systems (Aleven, McLaughlin, et al., 2016; Cronbach & Snow, 1977; Kalyuga, 2014; Kozma, 1991). Consider the following examples that explore just one of many relevant student characteristics that affect learning outcomes. Students with higher levels of knowledge require less instructional support than low knowledge students (Kalyuga, 2007a). Novice students who saw still graphic images learned more than text alone and animated graphics. Experienced learners students showed no differences in learning between the three formats (ChanLin, 2001). Observing deep level reasoning questions in tutoring improves
learning for low prior knowledge students more than high prior knowledge students (Craig et al., 2012).

ITS design is partially motivated by ATI research. One of the four main components of ITSs is the student model, which tracks the student’s knowledge state and other qualities pertinent to learning. There is a long line of research that explores how intelligent tutoring systems should best adapt to the student’s knowledge state (Aleven, Mclaughlin, et al., 2016). However, there remains a need for research that directly compares and carefully documents how learner prior knowledge may interact with specific design features in conversation-based ITSs. ITS designers include many different well-documented pedagogical and multimedia learning strategies in their systems (see the Design Recommendations for Intelligent Tutoring Systems series; Sottilare et al., 2013, 2014, 2015, 2016). These strategies are assumed to be effective, and often are. However, the various pedagogical strategies are not always carefully compared to alternate strategies, controls, or with potential ATIs. There is a need for ablation studies that carefully turn on or turn off individual system features and observe its effect on learning. For example, researchers have suggested that within conversation-based ITSs, vicarious learning and learning in triologues with student agents are best for low-knowledge students, whereas teachable peer agents are best for high-knowledge students (Graesser, 2016; Graesser, Forsyth, et al., 2017; Millis et al., 2011). Low prior knowledge students are expected to perform best when a peer agent is available, so they can vicariously model the peer agent’s productive learning behaviors. Redirecting negative feedback to peer agents is assumed to prevent students from becoming discouraged. Another assumption is that adaptive hints and prompts provided by the tutor agent will best support medium to high level students because they increase learner interactivity and engagement. While there is some data that supports these claims, it is sparse and underwhelming.
One goal of this dissertation is to fill this research gap by directly comparing the effectiveness of different conversational frameworks and determining how they interact with learner domain knowledge in AutoTutor (Graesser et al., 2003; Graesser, Chipman, et al., 2005; Graesser et al., 2010, 2014; Nye et al., 2014). AutoTutor is a conversational ITS that can implement various learning or conversational frameworks (e.g., vicarious learning, dialogues, triologues). In the proposed experiment, the content will be delivered to students via AutoTutor in either an interactive dialogue setting (i.e., student interacts with a tutor agent) or a vicarious setting (i.e., student observes a recording of an interactive dialogue). The results of this study will help inform the future design implementations of ITSs; particularly how conversational frameworks can adapt to the learner given their level of domain knowledge.
Chapter 2. Human Tutoring: Theory and Research

Learning in Tutoring Theory

The pedagogical strategies of tutoring are naturally grounded in the more general cognitive theories of learning. Often, the research of human tutoring and ITS efficacy tacitly assumes the applied tutoring strategies fit neatly into the well-established theoretical framework for learning. If certain tutoring strategies frequently fail to promote learning the question should be asked if the strategy is well-aligned with learning theory, or if there are specific circumstances (e.g., unique learner or domain characteristics) when traditional learning strategies simply do not apply. For example, it is well-established that students who are tutored receive larger learning gains than students in a traditional classroom setting. Specifically, tutees receive an average improvement ranging from 0.4 to 2.3σ over students in a traditional classroom setting and other controls (Bloom, 1984; Cohen et al., 1982). In this setting students are more engaged and active in constructive learning, in contrast to traditional classrooms where students are passive receivers of knowledge (Lave, 1990). A common distinction between passive and interactive learning is drawn from constructivist theories.

The constructivist model proposes that some active learning behaviors are critical for developing an accurate and deep understanding of new knowledge. In this model, knowledge is only constructed within the learner while the teacher is only indirectly involved in this process (Prawat, 1996). Constructivism has been described from several perspectives and has a long history of shaping pedagogy across a wide range of domains (Graesser et al., 1995; Jonassen et al., 1999; Von Glasersfeld, 1991). Students construct new knowledge by making inferences, elaborating, and integrating new materials (M. T. H. Chi et al., 2001). Additionally, knowledge construction can be supported through forming hypotheses, reflecting, and summarizing the
learning content (Chan et al., 1992; Palinscar & Brown, 1984). Tutoring pedagogy grounded in constructivism should then optimize opportunities to promote the above behaviors.

Compared to passive students, active students achieve deeper learning and levels of comprehension than passive students (Brown & Palincsar, 1989). Ideally, tutors provide opportunities for students to construct knowledge of deep-level concepts. The concept of deep learning appears frequently in the learning literature (Bloom, 1984; Craig et al., 2006, 2012; Dzikovska et al., 2014; Gholson et al., 2009; Graesser, McNamara, et al., 2005; Graesser et al., 2010; Graesser & Person, 1994). Deep-level concepts or deep-level learning is based on idea that comprehension and learning exists on a spectrum that ranges from shallow to deep. Bloom’s Taxonomy of Educational Objectives (Bloom et al., 1984) was originally developed in the 1950s and organized the different levels of comprehension into a cumulative hierarchy consisting of six major categories. Brief descriptions of the six levels of comprehension are provided below.

**Knowledge** – Comprehension at this level involves the simple memorization of previously learned material without understanding the meaning of the material.

**Comprehension** – Comprehension is the ability to grasp the meaning of the learned material. Comprehension can be displayed through interpretation, translation, and extrapolation of the learned material.

**Application** – Comprehension at this level involves the application of learned material into new situations

**Analysis** – Comprehension at this level involves drawing connections, organizing, and understanding relationships between individual components of a concept.

**Synthesis** – Synthesis is the ability to form the individual components of a concept into a new whole. Synthesis can be displayed through producing unique communication, plans, or sets of operation, as well as deriving a set of abstract relationship between concepts.

**Evaluation** – The highest level of comprehension involves the ability to judge the value of learned material for a specific case. At this level, the learned material can be evaluated in terms of its internal evidence or external criteria.
If a student leaves a tutoring session with a deep-level understanding of the content they are better equipped to transfer their knowledge into future lessons (Craig, et al., 2000; Craig et al., 2006). However, there are concepts that are beyond reach for students at any given moment for a variety of reasons, including a lack of prerequisite knowledge or necessary guidance.

Vygotsky describes the “Zone of Proximal Development” (ZPD) as the distance between an individual’s level or capacity to currently solve a problem and the potential level to solve a problem with “adult guidance” or peer help (Chalkin, 2003; Vygotsky, 1987). In other words, a student is in the ZPD if they are learning content that is not too simple that they could solve it without guidance but is not too difficult that they could not solve it with guidance. The concept of ZPD is routinely applied in problem-based educational settings. However, striking a balance between the promotion of deep-level learning while also remaining in a student’s ZPD can be difficult but is thought to be more easily achieved in learning environments that can personally adapt to the individual such as tutoring (Harland, 2003; Wood & Wood, 1996). A key goal of ITSs is to automatically and intelligently adapt to students to keep them within their ZPD (du Boulay & Luckin, 2001; Graesser et al., 2001; Murray & Arroyo, 2002; Woolf, 2009), which is described in more detail in Chapter 3.

**Human Tutoring Strategies & Effectiveness**

Bloom (1984) proposed “The 2 Sigma Problem” in a pivotal paper for research in tutoring and ITS design. In six different studies, Bloom and his graduate students observed that students who were tutored by a “good” human tutor performed two standard deviations higher than students in conventional classroom controls. That is, the average learner in the tutoring condition achieved higher learning than 98% of the students in the classroom condition. Bloom noted that teachers in conventional classroom settings would direct their teaching and feedback to some
students but ignored others. Bloom (1984) further observed the restrictions of the conventional classroom setting:

Teachers are frequently unaware of the fact that they are providing more favorable conditions of learning for some students than they are for other students. Generally, they are under the impression that all students in their classes are given equality of opportunity for learning (pp. 11).

Bloom points out that while one-on-one tutoring appears to be vastly superior to traditional classroom teaching, it is too expensive and impractical for widespread implementation. He refers to this as the “2 Sigma Problem.” Bloom (1984) reviewed six different studies that attempted to solve the 2 Sigma Problem by comparing different classroom instructional strategies as well as different combinations of these strategies in the classroom (e.g., reinforcement, cues and explanations, student time on task). In total, six different solutions fell short of the $2\sigma$ effect size seen in one-on-one tutoring.

High tutoring efficacy is frequently reported in other research, but with a more modest effect size. For example, a more recent meta-analysis reported the average effect size of human tutoring to be $0.79\sigma$ when compared to classroom teaching (VanLehn, 2011). Cohen et al. (1982) performed a meta-analysis which reviewed a large sample of studies comparing human tutoring to traditional classroom instruction. Most of the tutors in these studies were not professional or expert tutors, and instead were peer tutors. Interestingly, the average learning gain remained 0.4 standard deviations above reading controls and standard classroom instruction. This suggests that regardless of the tutor’s experience there is something inherent in the tutoring process that underlies its effectiveness. Similar results were observed by (Muldner et al., 2014) in a vicarious setting. Although students were unable to directly interact with a tutor, their results aligned with previous research; nonexpert tutors can sufficiently provide the benefits of tutoring.
Peer tutoring, both interactive and vicarious, presents a possible solution to the affordability issue of the 2-sigma problem. While the metanalyses following Bloom’s work (1984) indicate the effect size of tutoring falls short of 2σ, Bloom (1984) sets the bar and challenges learning researchers to develop tutoring strategies to solve the 2-Sigma problem. One-on-one tutoring routinely outperforms traditional classroom teaching. The disparity between the two settings warrants a closer examination of tutor and student behaviors during tutoring.

One approach for exploring tutoring effectiveness is found in the collaborative theory of communication, which asserts that the student’s role in the dialogue with the tutor is critical for understanding the content (Schober & Clark, 1989). Participants of a dialogue establish common ground and an understanding of each other through collaboration. Generally, tutors follow a 5-step conversation framework (Graesser et al., 1995; Graesser & Person, 1994; Person et al., 1995). Below are the five steps described in Graesser et al. (1995, p. 504):

   Step 1: Tutor asks question
   Step 2: Student answers question
   Step 3: Tutor gives short feedback on the quality of the answer
   Step 4: Tutor and student collaboratively improve the quality of the answer.
   Step 5: Tutor assesses student’s understanding of the answer.

The tutoring components and strategies below can all fit into this conversational framework. Specifically, question asking, explanations, and scaffolding are components of the above five-step tutorial framework. Closer examinations of these components provide insight into the origin of tutoring effectiveness.

Chi et al. (2001) reviewed different investigative approaches to understanding the effectiveness of tutoring and placed them into three categories: a tutor-centered hypothesis, a student-centered hypothesis, and an interactive hypothesis. The tutor-centered hypothesis (T-hypothesis) posits that the effectiveness of tutoring derives from the tutor’s behaviors and
strategies. The student-centered hypothesis (S-hypothesis) suggests that tutoring effectiveness derives from the student’s efforts and self-generated responses to tutor actions. Finally, the interactive hypothesis (I-hypothesis) is the idea that tutoring effectiveness comes from the advantages provided by interaction opportunities (e.g., interactive dialogues). Dialogues can contain interactive and non-interactive content (M. T. H. Chi et al., 2001). In interactive dialogues, tutor responses and comments request information from the student. Whereas in non-interactive dialogues tutor responses are simply explanations of concepts that do not provide opportunities for students to respond. Chi et al. (2001) conducted two studies that compared different types of tutor and student moves to deep-level and shallow-level learning during a tutoring session. The study involved eleven tutors who were college students and novice tutors with expertise in the domain material, the human circulatory system. Each tutor interacted with one of eleven 8th grade students during a single tutoring session. Transcripts of the tutoring sessions were coded into various categories, including tutor statements (e.g., giving explanations, answering questions, scaffolding with prompts) student statements (e.g., unprompted self-explanations, asking questions, answering questions). Their findings indicate that certain statements/moves associated with all three hypotheses were correlated with learning. For the T-hypothesis, results of a stepwise regression indicate the “giving explanations” move by tutors significantly improved the accuracy of model ($\Delta R^2 = .115, p = .009$), but notably only for shallow learning. The S-hypothesis was supported in that two student moves were correlated with learning. “Student responses to scaffolding” significantly improved the model ($\Delta R^2 = .280, p = .01$) whereas “student reflective comments” was correlated with deep learning ($\Delta R^2 = .261, p < .001$). The I-hypothesis was supported in that elicited student responses were correlated with learning, but self-initiated student reflections were not.
The above study further indicates that many features of tutoring can be effective at promoting learning and perhaps counterintuitively, not all these features involve interaction. Tutor behaviors, student behaviors, and interactive behaviors may affect deep level learning and shallow level learning differently. Interestingly, Chi (2001) found that the only variable associated with deep level learning, outside of student prior knowledge, was “student reflective comments.” Many tutor moves, student moves, and interactive moves thought to be effective did not contribute a significant amount to the learning in these sessions. A similar trend was reported in (VanLehn, 2011), who reviewed the general effectiveness of tutoring as well as various tutoring features traditionally thought to support its effectiveness. For example, one intuitive hypothesis is that human tutoring is effective because tutors can detect and diagnose student misconceptions. However, several studies indicate student misconceptions often go unnoticed and generally tutors are not accurate judges of student knowledge (Begeny et al., 2011; M. T. H. Chi et al., 2004; Putnam, 1987).

**Question Asking in Tutoring**

One of the more obvious distinctions between classroom-based teaching and tutoring is that tutoring often involves an ongoing dialogue between the learner and tutor. Naturally, question asking is a core component of the tutorial dialogues. For example, students ask clarification questions and information seeking questions whereas tutors provide guiding questions and questions to assess the student’s current understanding. Question asking is a major component of the tutorial dialogue and appears in most of the Graesser et al. (1995) five-step conversation framework (i.e., Step 1, Step 4, and Step 5). In general, student retention is improved when they are asked questions that induce deep levels of encoding compared to questions that only induce shallow levels of encoding (Craik & Tulving, 1975). Generating questions is considered a
fundamental component in cognitive processing that supports comprehension and is associated with learning complex material and problem solving (Graesser et al., 1996; Palinscar & Brown, 1984). Students, on average, are not skilled at asking good questions. However, if they are taught to ask good questions that seek out deeper levels of comprehension such as deep-level reasoning questions (Bloom et al., 1984; Graesser & Person, 1994) they improve their comprehension, learning, and recall of technical material (Craig et al., 2000; Davey & McBride, 1986; Driscoll et al., 2003; Palinscar & Brown, 1984).

Graesser and Person (1994) observed one-on-one human tutoring sessions of college students and 7th grade students. They focused on question asking behaviors of both the tutors and students during the tutoring session. The questions were classified into 20 different categories based on the (Graesser et al., 1992) question asking taxonomy. Additionally, the questions were classified on depth by using Bloom’s “taxonomy of educational objectives” which sorted educational tasks based on cognitive depth of comprehension (Bloom et al., 1984). Shallow questions included questions that repeated tutor questions and questions for clarifying something the tutor said. Deep-reasoning questions included: causal antecedent questions (What caused an event?), causal consequence (What are the consequences of an event?), goal orientation (What are the motives behind an agent’s actions?), instrumental/procedural (What instrument or plan allowed an agent to accomplish a goal?), and enablement (What object or resource allowed an agent to perform an action?). They found majority of student questions were shallow questions; 70% of student questions were considered shallow (level 1 of Bloom’s educational objectives) and only 30% of the questions spanned levels 2 through 6 of Bloom’s taxonomy. However, they did observe a positive correlation between student achievement and deep questions. For tutors, 60% of their questions fell into the short-answer category, with the most frequent category of being
verification questions (43%; Yes/No answers) that focused on determining student knowledge deficits. Only 35% were open-ended questions that gauged a student’s understanding at a deeper level. This question category included definition, comparison, example questions, interpretational, judgmental, antecedent, consequence, goal orientation, enablement, instrumental/procedural, and expectational questions. The remaining 5% of tutor questions fell into the “request/directive” category.

Unfortunately, tutors have difficulty determining student misconceptions (Begeny et al., 2011; M. T. H. Chi et al., 2004; Putnam, 1987). A contributing factor is that poor-performing students are not good at judging their own understanding of a topic and will often respond with a “yes” when asked if they understand (M. T. H. Chi et al., 1989). Adding to the problem, just as students have a poor understanding of their own misconceptions, tutors rarely ask questions that can target student misconceptions (McArthur et al., 1990; Putnam, 1987). If tutors are provided information about the student’s overall mastery level, tutors changed their behavior, improving effectiveness (Wittwer et al., 2010), but tutors are not more effective when they are given more detailed information about their tutee’s specific errors and misconceptions (Sleeman et al., 1989). However, an approximate knowledge of a student’s understanding may be enough to push the tutorial dialogue forward, ultimately promoting learning. The above details an obvious limitation of average human tutors and highlights an area where ITSs can improve on human tutoring by considering detailed information about misconceptions and adjusting their tutoring moving forward.

Self-Explanations

In tutoring, the tutor works with the student to help construct accurate explanations in response to questions asked by the tutor. The tutor will ask the student a question, wait for an
answer, and then work with the student to construct an accurate and complete explanation to the question. In this setting, students direct their explanations to the tutor and the tutor can correct any misconceptions in the explanation. Through scaffolding, the student eventually integrates the new knowledge into their pre-existing knowledge structure. Self-explanations, on the other hand, is another strategy that promotes the integration of new information with existing knowledge. Self-explanation is an interactive learning behavior in which students generate explanations to themselves when studying a text or working through a problem (M. T. H. Chi et al., 1994). Producing self-explanations involves several cognitive mechanisms, which are also often found in tutoring. These include: “…generating inferences to fill in missing information, integrating information within the study materials, integrating new information with prior knowledge, and monitoring and repairing faulty knowledge” (Roy & Chi, 2005). The increased depth of processing required during self-explanations is also related to improved learning and learning of more complex material.

The “self-explanation effect” describes a general finding that students who are prompted to self-explain in a variety of tasks (e.g., when solving problems, reading texts, completing incomplete figures) regularly outperform students in control conditions (e.g., reading a text twice, speaking aloud) (M. T. H. Chi, 1996; M. T. H. Chi & Wylie, 2014). For example, Chi et al. (1994) conducted a pre-test post-test experiment comparing learning between students who were prompted to self-explain and students who were not prompted to self-explain. Students in the self-explanation condition showed significantly more learning that those in the control. Also, the difference in learning gains between the two groups was larger for questions that were more difficult (M. T. H. Chi et al., 1994). Additionally, students who spontaneously elicit self-explanations tend to perform better than students who do not (M. T. H. Chi et al., 1989).
Prompting students to self-explain consistently promotes learning across a variety of domains. Self-explanation has been used to promote learning in a complex game environment (Clark et al., 2016), reading strategy training (McNamara, 2017), and computer skill acquisition (T.-Y. Chi et al., 2017). Note that each of these studies implemented self-explanation teaching strategies in a digital environment. A meta-analysis on self-explanation in mathematics indicates that self-explanation consistently leads to improved procedural knowledge, but performs best when used in conjunction with domain-specific scaffolding (Rittle-Johnson et al., 2017).

Strictly speaking, self-explanations are not directed to other agents (e.g., tutors, peers) and instead are self-directed. While self-explanations tend to exist outside of tutoring, this strategy can still be promoted during a tutoring session. In this case, self-explanations can arise during Step 4 (Tutor Improves Quality of Answers) of the common dialogue framework (Graesser et al., 1995). When a tutor prompts a student to provide a self-explanation, the student self-explains in the sense that they think through their own understanding of a knowledge component and visibly work through it in front of the tutor. There is some evidence that generating answers and explanations to tutor questions provides similar beneficial effects as self-explanations (M. T. H. Chi, 1996; M. T. H. Chi et al., 1989, 1994; Pirolli & Recker, 1994; Renkl, 1997). Tutoring provides more opportunities for students to interact with their learning environment which includes more opportunities to self-explain.

Tutor actions can be broken down into two different categories, prompting and scaffolding (M. T. H. Chi, 1996). Prompting, as described above in Graesser et al. (1995), can be applied without having any knowledge of the content domain. This includes questions like, “What do you think should happen next?” or “What does this mean?”. These questions encourage students to provide self-explanations. Tutors need a domain understanding to provide scaffolding.
Scaffolding steps include: orienting the student by describing the problem, providing the student a goal, and completing a student’s reasoning step (M. T. H. Chi, 1996; Graesser et al., 1995). The effectiveness of tutoring may come from the interplay of prompting students to provide self-explanations (self-construction of knowledge) and providing scaffolding from the tutor (co-construction). This deemphasizes the direct effect of tutor skill on tutoring outcomes and may provide an explanation for the results of Cohen et al. (1982); inexperienced peer tutors can still provide a highly effective learning environment.

It is important to note that self-explanations may not be accurate. Part of the self-explanation effect involves making accurate self-monitoring statements (“OK. I think I understand that”). (VanLehn et al., 1992). “Good solvers”, those who scored highest on a quantitative problem-solving task, used more self-explanations than poor solvers, but also had more accurate self-monitoring statements. Detecting and correcting inaccurate self-explanations is an important task for both tutors and ITSs. Self-explanation is a powerful pedagogical technique and discoursed-based learning environments like tutoring and ITSs should make an effort to include self-explanation and also a means to provide consistent and accurate feedback to correct inaccurate self-explanations.

**Vicarious Learning in Tutoring**

Part of the self-explanation effect is thought to be due to its active learning nature, however there is evidence that the self-explanation effect also applies in vicarious learning environments (Craig et al., 2012; Rummel & Spada, 2005). For example, Craig et al. (2012) included self-explanations in a vicarious learning environment by having a tutee agent provide self-explanations to content statements provided by a tutor agent. They found that the low knowledge learners performed best in conditions that included self-explanations when compared to question
only and monologue conditions. Additionally, asking good deep-level reasoning questions prompts students to apply deeper levels of encoding new information. Tutoring also provides opportunities for the student to ask questions. So far, the common theme for tutoring strategies is that they are all interactive in nature and encourage students to be active in their learning experience. However, vicarious learning appears to be another effective way to promote learning. Vicarious learning is knowledge acquisition that occurs when students are not addressed and simply observe the learning content without directly interacting with it. Vicarious learning stems from Bandura’s (1978) social learning theory and is often used interchangeably with observational learning (Bandura, 1978; Lee et al., 1998; Rosenthal & Zimmerman, 1978). Social learning theory began with Bandura’s early work in which children observed adults interact with a “bobo doll” aggressively (Bandura, 1978; Bandura et al., 1961, 1963). Bandura observed that children would then model the observed aggressive behavior and would also interact with the bobo doll in an aggressive manner. Central to the social learning theory is that learning is a cognitive process which occurs in a social context, and that individuals can learn from watching others. For example, Fox Tree (1999) found that students who overheard dialogues performed better in a referential communication task than students who overheard monologues. The general tutoring framework is powerful enough to provide benefits to outside observers, especially compared to monologue observers (M. T. H. Chi et al., 2017; Craig et al, 2000; Craig et al., 2009; Driscoll et al., 2003).

Students ask few questions during learning and when they do ask questions they tend to be at a shallow level, such as fact-seeking or clarification questions. As discussed in the question asking section above, generating questions is associated with comprehension and learning of complex material (Graesser et al., 1996; Palinscar & Brown, 1984). Teaching students to ask
good questions should benefit their comprehension of the immediate learning materials and ideally the question-asking behaviors will continue to help them with future learning. Craig et al. (2000) explored whether students would generate more questions on during a free-recall task in which they interacted with an experimenter who would answer any of their questions. The participants were divided into two conditions, a vicarious dialogue condition and a monologue condition. In the vicarious dialogue condition, the virtual tutee would ask deep-level reasoning questions, per the Graesser & Person (1994) taxonomy, the tutor would respond with an answer, and the tutor agent and tutee agent would carry out a dialogue. In the vicarious monologue condition, the tutor would respond to tutee broad questions (i.e., not deep-level questions) but in a monologue discourse and without a discussion between the two agents. Participants in the vicarious dialogue condition significantly outperformed participants in the vicarious monologue condition on an immediate learning task with a Cohen’s $d$ of 0.44. On a transfer task, participants were able to ask the experimenter questions while they were presented with eight new topics. The experimenter would answer each question until the participant finished the topic. Participants in the vicarious dialogue condition provided significantly more questions than those in the monologue condition. Additionally, participants in the vicarious dialogue condition generated significantly more deep-level reasoning questions than those in the monologue condition. Also, of the questions asked by participants in the monologue condition, a significantly greater proportion were shallow-level questions than the questions asked by participants in the dialogue condition.

The ICAP Framework

The ICAP framework (interactive > constructive > active > passive) proposes that students learn more as they become more engaged during learning (M. T. H. Chi et al., 2017, 2018; M. T.
H. Chi & Wylie, 2014). That is, students who are interactive during learning are more effective than students who are constructive, both behaviors are more effective than active learning behaviors, all of which are more effective than passive behaviors. This framework can be applied to learning activities beyond tutoring, but the interactive and constructive components are particularly relevant to “the gold standard” of learning environments. During tutoring, there are many opportunities for the tutor to encourage constructive behaviors while also interacting by through dialogue (i.e., tutor and tutee take turns providing their own insight, questions, and responses).

Chi and Wylie (2014) carefully operationalize each learning behavior in order to prevent any confusion with general terms or broader theories (e.g., constructivism), which also enables prior data generated from learning research to be interpreted within the ICAP framework. They operationally define passive learning as, “being oriented toward and receiving information from the instructional materials without overtly doing anything else related to learning” (Chi & Wylie, 2014, p. 221). An example of passive learning would be watching a lecture without taking notes or displaying any other learning behavior. Learner engagement is considered active if, “…some form of overt motoric action or physical manipulation is undertaken” (Chi & Wylie, 2014, p. 221). As opposed to passive learning, active learning during a lecture would involve the student taking verbatim notes or repeating the material in some way. They define constructive learning behaviors as, “...those in which learner generate or produce additional externalized outputs or products beyond what was provided in the learning materials” (Chi & Wylie, 2014, p. 222). In a lecture environment, this would involve the student reflecting on the learning material in some way, such as asking questions or providing a self-explanation. Finally, they define interactive learning behaviors as, “…dialogues that meet two criteria: (a) both partners’ utterances must be
primarily constructive, and (b) a sufficient degree of turn taking must occur” (Chi & Wylie, 2014, p. 223).

With the above definitions, the ICAP framework is supported by many different studies that encompass various domains and learning environments. The ICAP framework has important implications for tutoring. For example, not all dialogues are inherently interactive, some may only encourage constructive behaviors. If the dialogues predominate consist of tutor statements, then there are less opportunities for interactive learning behaviors. Alternatively, if the dialogues progress even though all student statements simply contain yes or no responses, then the tutoring is not interactive dialogues, per the ICAP definition.

At a first glance, the ICAP framework appears to discard vicarious learning environments as an inferior learning environment. However, students do benefit from observing and modeling beneficial learning behaviors of others, specifically other students who are being tutored (Craig et al., 2000; Fox Tree, 1999; Gholson, et al., 2009). Chi et al. (2017) acknowledge that while interactive environments are generally the most effective, they are also the most resource intensive and therefore less scalable and unavailable to most students. Chi et al. (2017) reviewed a series of studies that evaluated a new learning environment that may offer a solution to Bloom’s 2-Sigma problem. Specifically, students work in pairs as they observe a recording of a one-on-one tutoring session. The dyads of students work together to complete each task and answer each question posed by the tutor in the recording. This approach combines the benefits of vicarious learning with interactive learning. This format was evaluated by three studies which had similar findings (M. T. H. Chi et al., 2008; Craig et al., 2009; Muldner et al., 2014). The studies revealed that observing one-on-one tutoring in pairs significantly improve learning from pretest to posttest, and observing dialogues improves learning significantly more than observing
monologues. Additionally, the studies revealed that dyad observers of tutoring dialogues perform as well as those who were tutored directly, with no significant difference between the two settings.

Chi et al. (2017) took a closer look at tutor and tutee behaviors to determine if there are any tutee moves correlated with learning in the observer pairs. For tutees, they coded for incorrect comments or misconceptions, substantive comments, and questions. For tutors they coded for elaborative feedback and deep questions. Interestingly, they found that none of the five types of dialogue moves were significantly correlated with learning for the observer pairs. However, dialogue observers showed significantly more interactive and constructive behaviors than monologue observers. Chi et al. (2017) also saw that observers focused more on the tutee statements than the tutor statements. Additionally, while the monologue observers repeated and elaborated correct information from the tutor, they did not learn more than the dialogue observers who would repeat and elaborate on tutee statements, which were occasionally incorrect. Chi et al. (2017) suggested that the presence of the tutee allows observers to model positive learning behaviors, which they supported in their analyses. This result is aligned with previous studies of observers vicariously modeling good tutee behaviors (Craig et al., 2000; Twyford & Craig, 2017).

Chi et al. (2017) suggested that the presence of tutee “conflict episodes” is one explanation for why observers benefit from the presence of tutees. They described conflict episodes as those that occur when a tutee presents an incorrect response to a question and would then go through a cycle of correcting their error with the tutor. Previous research has also observed that the presence of conflicting statements can lead to confusion, which then leads to deeper learning as the dissonance is resolved (Lehman et al., 2013). Finally, the authors suggested that, like the
tutees, dialogue observers are novice students. Thus, the observers and tutees have what the authors refer to as a zone of representational match. They describe this representational match as both observers and tutees sharing a similar understanding and knowledge representation, both of which are unlike the knowledge representation of the tutor.

Students who participate in tutoring dialogue, or those who observe these dialogues, learn more than listening to monologues, like traditional classroom lectures. For example, dyads of students who observe one-on-one tutoring learn more than those who observe monologues (Chi et al., 2017). Several things drive this effect. First, tutoring involves interactive learning as opposed to passive learning and active students (e.g., those who form hypotheses and justifications, analogies, summarize and predict, and revise their knowledge) achieve deeper levels of comprehension than passive students (Brown & Palincsar, 1989; M. T. H. Chi et al., 2001). Also, according to the cognitive constructivism theory, students construct new knowledge by making inferences, elaborating, and integrating new materials (M. T. H. Chi et al., 2001). Tutoring provides opportunities for these constructive and interactive behaviors to occur. Secondly, tutoring involves more question asking, both from the student and tutor, than monologues (Graesser & Person, 1994). Questions activate key cognitive processes that are associated complex learning, problem solving, and deeper levels of comprehension (Graesser et al., 1996; Palinscar & Brown, 1984). Third, tutoring provides both prompting and scaffolding. Tutors can prompt students to provide self-explanations, which improves learning more so than students who are not encouraged to self-explain (Chi, 1996; Chi & Wylie, 2014). Students vicariously learn from watching a dialogue (Fox Tree, 1999), and model positive question-asking behaviors which is also associated with improved learning (Craig et al., 2000). When observing tutoring dialogues in pairs, students are more constructive and interactive and learn as much
students in one-on-one tutoring (Chi et al., 2017) and tutoring dialogues can be improved by including deep-level reasoning questions (Craig et al., 2006).

However, it remains unclear if tutoring can be improved by adjusting to the individual differences of the observers of tutoring. There must be some tutoring strategies that apply to students of all knowledge levels, given the findings reported above. ITS developers should consider how student differences interact with each system feature as they continue to create more scalable and individualized tutoring systems. For example, some students may only need to observe tutoring between a tutor agent and tutee agent, while others may need to interact directly with the tutor. Tutors often do not apply ideal pedagogical strategies, which suggests that there is room for improvement. ITSs offer a solution to these issues by providing a scalable training environment that can systematically apply ideal learning strategies that consider individual differences. The next chapter will review intelligent tutoring systems, their different designs, their effectiveness across various domains, and how they apply human tutoring research in a multimedia environment.
There is little question that tutoring is effective, but the average human tutor does not apply ideal teaching strategies and often does not accurately gauge their tutee’s understanding (Begeny et al., 2011; M. T. H. Chi et al., 2004; Putnam, 1987). Even if a tutor is aware of the many strategies that can improve their tutoring, they are likely unable to simultaneously keep track of their student’s understanding, know the best time to provide a deep-level reasoning question, when to prompt students to self-reflect, or when to provide more scaffolding. Additionally, human tutors select tasks based less on the student’s mastery or knowledge and more on a curriculum script (M. T. H. Chi et al., 2008; Graesser et al., 1995; Putnam, 1987). Human tutoring effectiveness is thought to be partially driven by more sophisticated tutoring strategies (e.g., Socratic irony), but tutors rarely utilize them (Cade et al., 2008; VanLehn, 2011; VanLehn et al., 2003). However, human tutors are very effective when they provide good feedback, good deep-level questions, scaffolding, and help motivate their tutees. Additionally, human tutors have a broad domain knowledge and can approach tutor misconceptions in a wide variety of ways, a behavior that ITSs struggle to replicate. One motivation for developing ITSs is that they provide an adaptive tutoring experience, improve on human tutoring by consistently applying effective tutoring strategies, while also avoiding common mistakes made by human tutors.

Learning and human tutoring theory provide the groundwork for ITS design, and they can be integrated into ITSs in any number of ways. ITSs provide a rich and controlled environment for rapidly exploring various interactions between specific tutoring strategies and individual differences. Research that meticulously explores how each minor feature of an ITS is resource

---

intensive, but necessary for improving the overall efficacy of ITSs and the efficiency of ITS design.

This dissertation is motivated by both observations of human tutoring and past ITS research. While interactive and adaptive ITSs are highly scalable, their development remains resource intensive. Nevertheless, as with traditional learning environments, higher doses of ITS use do not always result in better learning. Specifically, research suggests regular interaction with the system may not be helpful for some students. Other students benefit from moving quickly through the system, avoiding any extraneous cognitive burden. The proposed research will help determine when “less is more” by comparing learning between two fundamental ITS design features, interactive and vicarious learning.

Intelligent tutoring systems are electronic learning environments that intelligently adapt to users, implement effective pedagogical strategies, and sometimes resemble human tutoring (Shubeck et al., 2018). To implement effective teaching strategies in an ITS, developers must understand what makes one-on-one human tutoring effective. Chapter 2 established that tutoring allows for the following interactions to occur: tutors offer opportunities for students to self-explain, tutors can ask questions that encourage students to think more deeply on a topic, tutors can correct misconceptions paving the way for effective scaffolding, and students can ask and answer questions. Ideally, tutors cover content that is within a student’s ZPD. That is, they seek to cover content that a student does not fully understand and is not too difficult for the student to learn. In short, one-on-one tutoring appears to be fundamentally adaptive. For a tutoring system to be considered intelligent, they should, in some way, replicate this adaptivity. ITSs aim to extract and apply effective pedagogical strategies from tutoring (e.g., discourse between tutor and tutee).
ITSs are becoming easier to create. The tools for system developers interested in developing adaptive learning systems are becoming increasingly more available. For example, several robust text-to-speech APIs are now commonplace and readily available (e.g., Amazon Polly, Microsoft Text to Speech API, IBM Cloud Text-to-Speech). Further, it is becoming easier to interact with students in natural language with natural language processing toolkits (NLP). Some NLP toolkits are now free to use, allowing for even fledgling developers to create a natural language ITS. Aleven, McLaren, et al. (2016) illustrated that example-tracing ITSs can be developed without any programming skills using programs like Cognitive Tutor Authoring Tools and can still be effective at promoting learning gains with large effect sizes. ITS developers are also now more confident they can find an audience for their system. The potential user base continues to grow given the increased availability of mobile devices. Consider that as of 2017, 77% of U.S. adults now own smartphones and 51% own tablets (Pew Research Center, 2017). For U.S. teens, 73% have smartphones, and 87% have access to desktop or laptop computers (Lenhart, 2015). Given that it is easier to make ITSs and easier to find users, ITS developers face the challenge to responsibly develop their systems based on evidence-based pedagogical strategies.

Chapter 3 is divided into five sections. The first section covers common design features of ITSs and provide examples of different design approaches. The second section covers the overall effectiveness of ITSs. The third section focuses on common pedagogical strategies applied in AutoTutor. The fourth will cover section reviews vicarious learning strategies implemented in AutoTutor. The final section discusses the use of trialogues, a current trend in conversation-based ITSs, and their potential to inform the more traditional dialogue approach.
**Approaches**

Ideally, an ITS ties together the literature on best practices in tutoring, learning models from cognitive psychology and the learning sciences, and evidenced-based human-computer interaction design features. This is not a simple task and given the differences between each relevant field (e.g., HCI design, cognitive psychology, computer science, data science). Two well-designed ITSs can look very different. However, ITSs do share some common features. They have four major components: the domain model, the learner model, the tutor model (or pedagogical model), and the user interface (Sottilare et al., 2014; Woolf, 2009). The domain model attempts to target specific qualities in the learning material that merit consideration during instruction (e.g., procedural domains vs. conceptual domains). For example, ITSs that cover well-defined domains like mathematics may be more rigid in their assessment and feedback than ITSs that cover broader domains like scientific inquiry. The learner model details the various cognitive, affective, and motivational states of a student at any given point during their interaction with an ITS. The data stored in a learner model includes information about a student’s prior performance on similar problems, the time it takes a student to respond to a question, or the emotional valence of natural language input. The learner model can also utilize information about the student determined prior to instruction, like general aptitude, grade level, or prior domain knowledge.

The learner model is closely tied to the domain model in that the learner model tracks a student’s mastery of knowledge components (KCs) that are either essential to understand future content or belong to a subset of other similar KCs. The pedagogical model determines which actions and appropriate pedagogical strategies (e.g., self-explanations, hints, review) the tutor should apply next, given information from both the learner model and domain model. For
example, if a student takes several attempts to correctly answer a problem, the tutor may decide to provide an isomorph of that problem to ensure the KC is truly mastered. In some systems a student can guess the correct answer with a simple trial-and-error strategy or other “gaming the system” behaviors (Baker et al., 2004). Alternatively, if a student answers a problem correctly without help, the ITS may determine the KC has been mastered and decide to move on to a new KC. Finally, the user interface is used to take in and interpret student input via various channels (e.g., speech, typing, clicking, eye movement). Additionally, the user interface is used to display a variety of media to the student (e.g., text, videos, images, animated pedagogical agents).

The pedagogical model, domain model, student model, and user interface all fit into either the inner-loop or outer-loop of an ITS. VanLehn (2006) describes the outer loop as the ITS actions that determine which tasks should be selected next. For example, after a student solves a question that covers a particular KC, then the system should determine if the student should receive a problem on the same KC or move on to a different KC. He describes the inner loop as the ITS actions that occur during each step the student takes to solve a problem. Decisions made within the inner loop involve assessing student input and then providing corrective feedback, hints, or prompts.

There is a wide variety of approaches to implement the inner loop and outer loop, as well as the four central components of ITSs. For example, Cognitive Tutor (Aleven & Koedinger, 2002) is a problem-solving tutor that adapts the ACT-R theory of cognition to effectively teach algebra and geometry. A different ITS, MetaTutor promotes self-regulated learning strategies to teach students about the circulatory system (Azevedo et al., 2009). iSTART focuses on self-explanation techniques to teach students reading-comprehension strategies (McCarthey et al., 2017;
McNamara et al., 2004). Some ITSs promote learning in virtual worlds. For example, VCAEST (virtual civilian aeromedical evacuation sustainment training) is a system that situates students in a virtual city block that was recently struck by an earthquake (Foronda et al., 2016; Shubeck et al., 2016). In VCAEST, students interact with an AutoTutor-based ITS as they triage victims. Crystal Island is a game-based virtual world ITS that teaches students microbiology concepts (Rowe et al., 2011; Taub et al., 2017). Operation ARA is another game-based ITS that uses an AutoTutor style conversation framework to teach scientific reasoning concepts (Cai et al., 2011; Millis et al., 2017). These systems, and many others, utilize different technologies and pedagogical strategies, each of which are effective at promoting learning across in their targeted domains.

Effectiveness

ITSs generally produce learning gains that range from 0.4 to 1.1 standard deviations higher than traditional classroom environments (i.e., lecture-based instruction) and reading controls (J. A. Kulik & Fletcher, 2016; Ma et al., 2014; VanLehn, 2011). For example, VanLehn’s (2011) meta-analysis saw an ITS effect size of $d = 0.76$ when compared to static reading materials and other controls. In other words, VanLehn (2011) found that step-based ITSs were nearly as effective as average human tutors, but still less effective than expert human tutors. Similarly, Ma et al. (2014) saw positive effects for ITSs when compared to no reading ($g = .35$), non-ITS computer based instruction ($g = .57$) and large lecture based learning ($g = .42$), with no significant difference between ITS learning and individualized human tutoring or small-group instruction. Additionally, (J. A. Kulik & Fletcher, 2016) meta-analysis of ITSs reported that, on average, ITSs raised test scores 0.66 standard deviations over conventional levels. Specifically, of the 50 controlled experiments included in their analysis, 92% saw ITSs produce larger
learning gains than the controls. Comparatively, a meta-analysis on computer-assisted-instruction tutoring reported an average effect size of 0.31 (C. L. C. Kulik & Kulik, 1991).

Interestingly, Steenbergen-Hu and Cooper (2014) reported much more modest effect sizes of ITS on mathematics instruction for K-12 students (from $g = 0.01$ to $g = 0.09$) in their meta-analysis. Their analytical methods varied from previous meta-analyses and their study inclusion criterion was less stringent. For example, no restriction was placed on studies that compared two conditions (ITS vs control) that did not control for all differences between the groups, making it more difficult to determine the source of the learning effects. Several quasi-experimental studies were also included in the Steenbergen-Hu and Cooper (2014) analysis. While the study settings tended to be more ecologically valid (e.g., classroom settings over entire semesters), experimenters were unable to control for differences in teachers, teacher support, time on task, etc. However, some important trends can be gleaned from their analysis that are particularly relevant for the current study. For example, in a fixed-effect model, general to high-level achievers benefit more from ITSs than low achievers ($g = 0.04$, $g = -0.18$ respectively, $p = .002$), however the significance was lost in a mixed-effect model.

While the evidence may be weak, it is aligned with previous research on computer-supported classroom studies, which found that low achievers and high achievers use ITSs and computer-based instruction differently (Hativa, 1988; Hativa & Shorer, 1989; Wertheimer, 1990). Specifically, low achievers make more system-related mistakes than high achievers. They suspected high achievers are more capable at adjusting to new learning environments than low achieving students (Hativa, 1988). Steenbergen-Hu and Cooper (2014) suggested that ITSs function best when students have adequate prior knowledge and self-regulating skills, motivation, and computer-skills.
Ma et al. (2014) conducted a meta-analysis that also explored the overall effectiveness of ITSs compared to human tutoring \((g = .11)\), non-ITS computer-based instruction \((g = .57)\), traditional teacher led classrooms \((g = .42)\), textbooks \((g = .35)\), and small group instructions \((g = .05)\). They found no significant difference between ITS learning and human tutoring or small group instruction. Ma et al. (2014) reported the effect size for low, medium, and high prior knowledge students. The effect size of ITS for low prior knowledge \((g = .38)\) and medium students \((g = .28)\) were significant, but none of the prior knowledge groups were significantly different from each other. The authors did warn that these results should not be interpreted as conclusive given the small sample size of studies who reported prior knowledge, and even fewer studies included students with high prior knowledge. Their results run counter to the Steenbergen-Hu and Cooper (2014) suggestion that adequate prior knowledge is necessary to fully benefit from ITSs.

VanLehn’s (2011) analysis took a closer look at features of ITSs based on human-tutoring pedagogy thought to be driving the benefits of ITSs. VanLehn (2011) hypothesized that feedback of human tutoring helps students repair their knowledge, and that human tutors scaffold their students by pushing them through the lesson through collaboration and indicating when they should continue. Studies included in VanLehn’s meta-analysis compared 4 types of tutoring: human tutoring, substep-based ITS, step-based ITS, and answer-based ITS. VanLehn hypothesized that the granularity (e.g., specificity, detail, frequency) of the feedback and scaffolding would affect student learning. He reasoned that finer-grained feedback targets smaller lines of reasoning (i.e., smaller knowledge components), thus improving overall learning. In human tutoring, there is no restriction on the amount of detail that can be provided in each instance of feedback. In step-based ITSs, students provide input for each of the steps they take to
complete a problem, which a tutor evaluates and provides feedback accordingly. Substep-based tutoring systems can provide feedback on even more specific steps required to complete a problem. Answer-based systems only provide feedback given the student answer to an entire problem. Here, feedback typically comes in the form of “correct” or “incorrect” followed by a hint or a different problem example. The steps required to reach the answer are not taken into consideration in answer-based systems, which require students to cross larger gaps between feedback and the targeted knowledge component.

The interaction-granularity hypothesis was not supported. Tutoring effectiveness was not observed to steadily increase with more specific feedback. Instead, he observed a granularity-interaction plateau. Human tutoring is roughly as effective as substep-based tutoring, which is about equal to step-based tutoring, all of which are more effective than answer-based tutoring systems. Both step-based tutoring systems and human tutors appear to provide sufficient scaffolding and feedback, which allows students to provide correct responses to questions and solutions to problems. Although the granularity differs between substep, step-based, and human tutoring, the differences are negligible. VanLehn (2011) explained that these results align with the ICAP framework.

**AutoTutor**

The proposed experiment will use an AutoTutor-based ITS to teach basic electronics engineering concepts, or scientific reasoning skills, to college level students. AutoTutor (Graesser, 2016) is a conversation-based ITS. A tutor agent holds a natural language dialogue with a student. This means AutoTutor students respond to tutor agent questions by typing, and more recently speaking, their responses. AutoTutor has been implemented in many different systems across several domains, including physics, computer literacy, critical thinking, medical
tria.ge, breast cancer awareness, and electronics engineering. AutoTutor’s framework is consistent across these diverse domains. Regardless of the domain, each AutoTutor iteration involves an expectation-misconception tailored dialogue (EMT) framework, a design based on observations of tutor-student interactions in human tutoring (Graesser et al., 1995). AutoTutor’s design uses various strategies known to promote learning, including ideal pedagogical strategies that human tutors rarely enact. Given that AutoTutor’s design is domain-general and flexible, it also functions as an ideal research tool for investigating the nuances of student learning during tutoring.

**Authoring AutoTutor Scripts**

AutoTutor dialogues are created by AutoTutor script authors, and a series of rules set in the AutoTutor Conversation Engine (ACE; Cai et al., 2015; Graesser et al., 2012) determines when each dialogue move should be used given the student’s current knowledge state. Figure 1 depicts the architecture of the AutoTutor conversation engine (ACE).

![Figure 1. AutoTutor Conversation Engine. From Applied natural language processing and content analysis: Identification, investigation and resolution, P. M. McCarthy & C. Boonthum (Eds.). Graesser et al., 2012. AutoTutor. (pp. 169–187). Hershey, PA: IGI Global](image-url)
Authors determine how many hints are necessary to adequately cover the expectation content. Likewise, authors will determine how many prompts are needed if all hints are exhausted and the student is still missing a key term or phrase. Misconceptions are also authored by script editors and function in much the same way as expectations. Misconceptions are addressed by the feedback dialogue move. AutoTutor can also track meta-cognitive statements from the student (e.g., guess, confused, know, forgot, lost) and respond accordingly. Each AutoTutor script can be altered in one of two ways. Beginners are encouraged to use the online AutoTutor authoring tool. More experienced authors can edit the .xml file directly and then upload the file to the authoring tool. Additionally, AutoTutor authors can adjust the conversation rules (e.g., if the student answer does not meet the correct answer threshold for a question, provide a hint for expectation 1). The default conversation rules often do not require any adjustments for a typical AutoTutor module. Frozen expressions, or canned expressions, are tutor agent utterances like “OK”, “Great!”, “Not quite” that come before or after a question, hint, pump, prompt, or summary. The curriculum script includes ideal answers, expectations, misconceptions, hints, pumps, and prompts. All interactions in AutoTutor are collected in log files. The log analyzer examines these log files, which allows AutoTutor researchers to take a close look at how AutoTutor interacted with the student during the lesson. The User State is determined by their progress throughout the AutoTutor module (e.g., which current expectation they are working to answer fully).

AutoTutor uses natural language processing to interpret student input and respond appropriately. Importantly, the student input does not need to make direct one-to-one matching of a bag of words provided by AutoTutor authors. Instead, student input is assessed by their meaning, or semantic matching, to ideal answers and expectations. To achieve this, AutoTutor
uses latent semantic analysis (LSA). LSA is an application of the semantic network representation of knowledge. It matches words according to their semantic overlap with other words. AutoTutor compares student input from each dialogue “turn”, or individual step during the dialogue, and then matches the content to either an expectation or ideal answer that was predetermined by the script author. If all expectations in a lesson are sufficiently covered, AutoTutor will advance the lesson to a summary step, provide a summary, and then either end the lesson or move on to a different main question.

To help correct the shortcomings of LSA (e.g., differentiating from negatives and positives in text, synonyms, handling numbers) AutoTutor uses regular expressions (RegEx). RegEx allows AutoTutor authors to target specific phrases and numbers that may be tricky to detect through LSA alone, including possible misspellings. LSA and RegEx are both used to evaluate student answers and their overlap with target answers and misconceptions. The combination of LSA and regular expressions in AutoTutor was compared to two human expert raters based on 892 student answers (Cai et al., 2011). The RegEx and LSA combination had a similar correlation with the two human expert raters ($r = 0.667$) as they did with each other ($r = 0.686$). These results suggest that AutoTutor’s natural language processing (NLP) computational mode adequately assesses learner input throughout their dialogue.

**Tutoring Dialogue in AutoTutor**

As introduced above, AutoTutor uses an EMT dialogue. This means that for each lesson in AutoTutor there are pre-defined expectations and misconceptions. Ideally, subject matter experts provide the expectations and misconceptions, where the expectations include the “ideal answer” to the main question of the lesson. Ideal answers are made up of several key pieces of information, or expectations, which are needed to fully understand the lesson. These expectations consist of smaller knowledge components which include smaller bits of information necessary
for understanding a single expectation. Students can move on to the next main question when they indicate they know all the knowledge components that make up each expectation, and all the expectations that make up the ideal answer. Altogether, a student will need to provide about a paragraph of information to fully answer a main question. AutoTutor applies specific dialogue moves, based on observations of effective human tutoring, to elicit student answers to each main question. These dialogue moves are tutor agent, and in some cases peer agent, utterances that drive the natural language dialogue. They include feedback, pumps, hints, prompts, and summarizations.

Shubeck et al (2018 pp. 139), provide a description of some dialogue moves in AutoTutor found below:

*Feedback*: Short feedback by the tutor that is either positive (e.g., Great! Good answer), negative (That’s not quite right), or neutral (Uh-huh, Okay). The type of feedback provided is determined by matching student responses to either expectations or misconceptions.

*Pumps*: AutoTutor provides pumps to the student so they will expand on their previously provided input (Can you expand on that a little bit?). Pumps are used to encourage students to generate more content and think more deeply about the main question or problem. The student has a better opportunity to provide matches to expectations or misconceptions.

*Hints*: These are tutor utterances that are more specific than pumps and target either a specific expectation or misconception. If the student is talking about a certain topic (e.g., function of computer RAM) but has not covered the topic fully, the tutor will provide a hint to move the student closer to fully answering the question and covering a particular expectation (e.g., What do you think would happen if a computer didn’t have RAM?).

*Prompts*: Prompts are provided to students to help them articulate a missing key term. These are follow-up dialogue moves when the student did not correctly respond to the previous hint (e.g., RAM stands for random access what?).

*Summarization*: This occurs when all hints and prompts are exhausted for all the expectations associated with a main question or problem. AutoTutor provides a summary of a good answer to the question or problem. In essence, the summary is a well-articulated ideal answer.
These are some of the AutoTutor dialogue moves generated by AutoTutor, noting that their content must be provided by the authors of AutoTutor scripts. Additional dialogue moves are described in Graesser, Forsyth, and Lehman (2017). AutoTutor scaffolds students through a pump → hint → prompt → assertion cycle. Below is a brief example of a dialogue in AutoTutor:

**Tutor Agent** (Main Question): What is the mathematical relationship between the total current, I, and the three branch currents I1, I2, and I3?"

**Student** (MetaCog): I don’t know.

**Tutor Agent**: Let’s work through it together.

**Tutor Agent** (Hint): In the given circuit, three currents I1, I2 and I3 are the parallel branch currents. How can we calculate the total current I, in terms of these branch currents?

**Student** (Bad Answer): One third of the summation of these three currents.

**Tutor Agent** (Feedback): No, not quite.

**Tutor Agent** (Prompt): The Law of Conservation of Energy says that the total output current must equal the input current at point A. So, the current I and the sum of currents I1, I2, and I3 must be what?

**Student** (Good Answer): Equal.

**Tutor Agent** (Feedback): Good.

**Tutor Agent**: I is equal to the sum of the three currents.

**Tutor Agent** (Closing): You got it. Let's try a new topic

The above example is a simplified version of an AutoTutor conversation and only targets one expectation. Typically for each main question there are several expectations and after the student adequately covers one expectation the tutor will initiate a discussion about another expectation. After all expectations are covered, the tutor agent will provide a closing statement and will either end the lesson or will move on to a different main question.
AutoTutor Effectiveness

Nye, Graesser, and Hu (2014) reviewed 17 years of AutoTutor research. They reported the average learning gain from pretest to posttest for AutoTutor is .8σ higher than standard controls. Again, while this learning effectiveness is significantly smaller than Bloom’s 2σ, more recent meta-analyses report more modest human tutoring effect sizes. Ma et al. (2014) reported an average human tutoring effect size of .4σ and VanLehn (2011) reported an average effect size of .79σ over controls. AutoTutor’s effect size is comparable to or greater than human tutoring, a significant feat considering human tutoring is referred to as the gold standard of learning environments. Nye et al. (2014) also reported a series of AutoTutor ablation studies that removed individual features (e.g., no avatar or voice, only voice and no avatar) and compared the resulting effectiveness to the full AutoTutor system. Interestingly, the results indicate that the modality of content delivery in AutoTutor does not play a significant role in its overall effectiveness.

Nye et al. (2014) also reviewed AutoTutor studies that noted how prior achievement interacts with the effectiveness of specific AutoTutor features. High-performing students regularly benefit from AutoTutor dialogues compared to controls (e.g., reading textbooks) as well as in vicarious settings, but appear to take better advantage of interactive AutoTutor modes than low-performing students (Jackson et al., 2006). Alternatively, low prior knowledge students show greater learning gains than high prior knowledge students in vicarious AutoTutor modes, such as observing the tutor agent teach a peer agent (Craig et al., 2012).

As discussed in Chapter 2, observing human tutoring can be as effective as interacting with a human tutor (Chi et al., 2017; Craig et al., 2000; Fox Tree, 1999; Gholson & Craig, 2006; Gholson et al., 2009). Moreover, VanLehn (2011) did not find support for the interaction-granularity hypothesis, which reasoned that learning performance should improve as the sub-steps and feedback become more detailed and individually tailored to the problem. Together,
these results suggest that further exploration is needed to determine when interactive dialogues are needed and how this need may vary for students with different levels of prior knowledge. The following sections describe how several multimedia learning principles and pedagogical strategies are implemented within AutoTutor.

*AutoTutor and multimedia learning principles*

While ITSs differ in many ways, they are all situated within multimedia learning environments. Applying teaching strategies of expert human tutors is obviously a critical part of the success behind ITSs, but the unique multimedia component of ITSs should also be designed in a way that considers multimedia learning principles. Multimedia learning is described as learning that occurs through instruction with both words and images (Mayer, 2014). The cognitive theory of multimedia learning assumes multimedia instruction that acknowledges cognitive functions will be more effective than multimedia instruction that does not consider these functions. Multimedia instruction should be designed in accordance to working memory theory. Individuals process visual and auditory information in two distinct perceptual channels (i.e., verbal or visual; Baddeley, 2003). Additionally, according to the cognitive load theory, there is a limited capacity for processing within the verbal and visual channels (Sweller et al., 2003). The cognitive theory of multimedia learning consists of a set of evidenced-based multimedia learning principles, several of which are described below.
Mayer (2014) describes the persona effect, or agent effect, as the tendency for animated pedagogical agents and their engagements with students to have a small but positive effect on learning. A meta-analysis assessed the effect of animated pedagogical agents on engagement and learning across a variety of intelligent tutoring systems and domains (Schroeder et al., 2013). They found that, compared to systems without animated pedagogical agents, ITS with agents improved learning with an effect size of $g = .19$.

Mayer (2014) describes the personalization principle as the tendency for students to learn more deeply when speech used in multimedia environments is more conversational than formal. People tend to interact with computers in much the same way as the do with others (Reeves &
Thus, multimedia environments that use more informal language than formal language result in students attributing more social qualities to the learning environment. AutoTutor script authors are encouraged to use more informal language than formal language to better align with the personalization principle, which is thought to enable a better sense of social presence (Mayer, 2014).

Chapter 2 illustrated that feedback is an essential component of tutoring that allows students to have opportunities to reflect on their own knowledge. The feedback principle is described in (Johnson & Priest, 2014) as the tendency for novice students to benefit more from explanatory feedback than corrective feedback alone. That is, simply telling a student whether their input is correct or incorrect is not as effective as a principle-based explanation for why their input is correct or incorrect. Johnson and Priest (2014) reviewed eight studies that directly compared explanatory feedback to corrective feedback by examining their effect on learning (see Mayer & Johnson, 2010; experiments 1 and 2 in Moreno, 2004); far transfer and near transfer scores in (Moreno & Mayer, 1999; Moreno & Mayer, 2005) far transfer and near transfer scores in (Moreno & Duran, 2004). On average, explanatory feedback provided a positive effect on learning with a mean Cohen’s $d$ of 0.72. Additionally, the Moreno (2004) experiments, the Moreno and Mayer (2005) experiment 1, the Moreno and Mayer (1999) experiments, and the Moreno and Duran (2004) experiments saw students who received explanatory feedback outperformed those who received corrective feedback only. There was also a trend in these studies that students who received explanatory feedback had deeper learning than those who received corrective feedback alone.

Research has shown that the effectiveness of multimedia instructional formats, such as frequency of feedback, guidance, and interaction, changes as students gain expertise in a domain.
This is referred to as the expertise reversal effect (Kalyuga, 2007a, 2014; Reisslein et al., 2006; Schnottz, 2010; Sweller et al., 2003). Initially, novices benefit from detailed instruction and increased interactivity with multimedia learning material. However, as novices gain more experience and higher levels of knowledge, they benefit from less detailed instruction and less interactive multimedia environments. High level students who receive less instruction in a multimedia environment perform better than high level students who receive more instruction. For example, Kalyuga et al. (2001) compared the effectiveness of a computer-based interactive format that was based on multimedia principles (e.g., modality principle, signaling principle, worked examples principle) to an instructional format that with less guidance and explanations that did not adhere to many of the multimedia principles. They found that format which adhered to multimedia learning principles was very helpful to novice students compared to the less interactive format. However, they found that as the students became more experienced in the domain, their performance was better in the less interactive format than in the more interactive format.

Several other studies support the expertise reversal effect. Reisslein et al. (2006) found that novices benefited more from worked examples than from problem solving alone, but more experienced students benefited more from problem-solving without worked examples. Brunstein et al. (2009) found that algebra students learned most from minimal guidance as they became more practiced, but algebra students with less practice benefited more from explicit instruction. Blayney et al. (2010) also observed the expertise reversal effect for accounting students. Novice students benefited most from sequentially entering isolated elements of formula into multiple cells, but more experienced students benefited best from entering the entire formula into a single cell.
One explanation for this effect is found in the redundancy principle (Mayer & Johnson, 2008). Information that may benefit novice students becomes redundant for high level students. From the prospective of cognitive load theory, processing redundant information requires additional cognitive resources, leaving less resources available for learning activities that benefit high knowledge students. Likewise, (Mayer, 1989) notes that a student’s prior knowledge is a key factor in how they construct conceptual models for new material.

The expertise reversal effect highlights the importance of designing multimedia environments that consider varying levels of student knowledge. ITSs should track student knowledge and adapt their feedback and interaction detail as students gain more experience. If the expertise reversal effect holds true in the proposed experiment, high level students in the interactive dialogue condition are expected to outperform high level students the vicarious dialogue condition. In the interactive condition they are expected to receive explanatory feedback as needed. In the vicarious AutoTutor condition high level students have no control of the pace of the lessons. In the proposed experiment, the feedback will be equivalent as much as possible across all conditions, but the modality in which they are presented to the learner will vary. In the interactive condition, feedback will target the student’s input. As such, some explanatory feedback may not be provided to students who already understand the content. Alternatively, in the vicarious condition students will observe corrective feedback to a different student’s mistakes. The feedback will contain the same content as the interactive conditions but will not necessarily target the observer’s understanding. The results of the study should provide insight into whether the modality of the feedback affects learning, as well as if the targeting and potential redundancy of the feedback affects learning. The following section describes the different ways feedback is delivered to students in AutoTutor.
Vicarious Learning in AutoTutor.

The ITS pedagogical strategies discussed so far involve students interacting with the system in some way, be it clicking on the interface, answering multiple choice questions, or carrying on a conversation. These strategies have proved effective and provide a learning environment where students actively construct knowledge through interactions. However, in some cases it is appropriate to implement a vicarious learning environment. Vicarious learning is also used in combination with interactive learning environments. Vicarious learning sometimes precedes interactive learning in AutoTutor and can be useful for promoting learning of shallow knowledge concepts that do not require interaction. However, there is research that indicates that in some cases, pure vicarious learning in AutoTutor and other systems is as effective or more effective than interacting with AutoTutor.

As described in Chapter 2, considerable research indicates that vicarious learning is often effective, particularly when students are observing a tutoring dialogue (Craig et al., 2000; Craig et al., 2006; Fox Tree, 1999; Gholson & Craig, 2006). Observing tutoring has shown to be an effective strategy for simple information delivery in a wide variety of domains such as physics (Craig et al., 2012; Gholson et al., 2009), the circulatory system (Craig et al., 2008), computer literacy (Craig et al., 2006), and scientific reasoning (Millis et al., 2011). In AutoTutor, vicarious conditions showed greater learning gains than controls for middle school students who have low prior knowledge (Craig et al., 2006; Gholson et al., 2009). Additionally, several studies have shown that low knowledge students can benefit more from vicarious learning than interactive learning conditions (Craig et al., 2012; Gholson et al., 2009; Millis et al., 2011). Low knowledge level students benefit from modeling positive learning behaviors of other students. For example,
vicarious modeling of positive goal-orientation is associated with increased learning (Twyford & Craig, 2017).

Research exploring vicarious learning in AutoTutor has yielded mixed results. For example, Craig et al. (2004) compared learning gains of low prior knowledge learners between individual vicarious observation of previously recorded student interactions with AutoTutor, collaboratively observing the recording, or interacting directly with AutoTutor. They found that students in the interactive condition learned more than both vicarious conditions, which aligns with the ICAP framework. However, Craig et al. (2006) saw vicarious students who observed a computer student interact with a computer agent while asking deep-level reasoning questions outperform students who interacted directly with AutoTutor. Craig et al. (2006) explored whether the difference in immediate learning between dialogue observers and monologue observers was due to the dialogue per se, versus the presence of deep-level reasoning questions. In their first experiment, they compared pretest-posttest learning gains of low prior knowledge learners between five conditions. In the first condition, interactive, participants interacted with a computer tutoring system that asked questions, responded to user input and provided hints. In the second condition, yoked-vicarious, participants viewed a recording of an interaction between a student and the computer tutoring system. In the third condition, monologue vicarious, participants viewed a monologue from a computer agent on the same materials, but the recording did not include a second agent. The fourth condition, half-questions-vicarious, participants viewed a recording of a computer agent providing content, deep-level reasoning questions, and then ideal answers. In this condition, the content was in a dialogue format, with a second agent asking deep-level reasoning questions prior to the delivery of an ideal answer from the tutor agent. In this condition, the expectation sentences, or information that a learner is expected to
know before completely understanding a topic, were not preceded by questions. In the final condition, *full-questions-vicarious*, the content was delivered in the same format as the half-questions-vicarious condition, but the expectation sentences were preceded by questions in a dialogue manner. The results of the first experiment saw participants in the full-questions-vicarious condition significantly outperform those in the other four conditions with an effect size of $d = 1.99$.

In a second experiment, Craig et al. (2006) compared pretest-posttest learning gains of low prior knowledge students between four conditions. An interactive condition and a yoked-vicarious condition were included and had the same format as in the first experiment. Two conditions, *deep-level reasoning questions monologue* and *deep-level reasoning questions dialogue* were also included. In the deep-level reasoning questions monologue condition participants saw a recording of a tutor agent who asked deep-level reasoning questions both before ideal answers were provided and before expectations were provided. In this condition, the same agent provided both the questions and answers, resulting in a monologue format. Alternatively, in the deep-level-reasoning questions dialogue condition, the questions were asked by a second, off screen, agent and the agent on screen provided the answers, resulting in a dialogue format. The results of the second experiment saw the deep-level-reasoning questions monologue condition ($d = 1.95$) and the deep-level reasoning questions dialogue condition ($d = 2.29$) significantly outperform the interactive and yoked-vicarious condition.

The results of Craig et al. (2006) highlight the importance of deep-level reasoning questions in dialogues during vicarious learning. Interestingly, the vicarious deep-level reasoning conditions outperformed interactive dialogue conditions in both studies, contradicting the “dialogue per se” hypothesis. However, it can be argued that each vicarious condition also
involved dialogue in some form, that is, students may have still interpreted the monologue conditions as dialogues because they contained questions with responses. Additionally, each condition showed significant learning from pretest to posttest with effect sizes comparable to or approaching 2.0, as seen in expert human tutors in (Bloom, 1984). The key take-aways from (Craig et al., 2000; Craig et al., 2006) is that students appear to learn from other students interacting with a tutor, specifically to ask better questions, and that including deep-level reasoning questions in the observed tutoring sessions improves their learning. These results are consistent with previous research who saw “overhearers” of dialogues perform better than overhearers of monologues on an instruction matching task (Craig et al., 2009; Fox Tree, 1999). Similarly, individual observers of dialogues outperformed those who observe monologue tutoring in multimedia learning environments, including observing animated computer agents (Craig et al., 2004; Driscoll et al., 2003; Muller et al., 2008). Further, within the collaborative observing paradigm, collaborative dyads who observe dialogues learn more than collaborative observers of monologues (M. T. H. Chi et al., 2017).

Using the same approach as their previous studies, Craig et al. (2012) evaluated the effect of virtual tutee explanations on learning and how this may interact with the student’s prior knowledge. In their first experiment, they compared learning gains of physics under four different conditions, each using animated pedagogical agents. In the monologue condition, a virtual tutor spoke 50 content statements while a virtual tutee listened. In the question condition (Q), a virtual tutee asked deep questions before each of the content statements were presented by the virtual tutor. In the explanation condition (E), the virtual tutor presented the content statements, and the virtual tutee provided an explanation by elaborating on the content statement. In the questions and explanations condition (Q + E), the tutee would ask the same
questions in the question condition which were followed by the same content statements from the tutor, which were then followed by with tutee explanations. Their results revealed a significant interaction between condition and knowledge level. Low prior knowledge students in the Q + E condition showed significantly greater learning gains from pretest to posttest (31% gain) than high prior knowledge students (7% gain). Further, the low prior knowledge students had slightly higher posttest scores than the high prior knowledge students. Learning gains for low prior knowledge students were compared between each condition and no significant differences were found, although there was a slightly higher gain in the Q + E condition approaching significance \( p = 0.08 \). High knowledge students in the Q + E condition showed significantly lower learning gains (7% gain) than those in the monologue condition (19%) and the Q condition (21%).

These results are aligned with previous research on the expertise reversal effect (Blayney et al., 2010; Brunstein et al., 2009; Kalyuga et al., 2001; Reisslein et al., 2006; Schnitz, 2010; Sweller et al., 2003). High knowledge students or high expertise students are disadvantaged when too much feedback, guidance, and instructional content are presented, presumably because the added content may not be aligned with their existing mental model. The cognitive load of high knowledge students increases when they receive redundant information (Kalyuga, 2014; Mayer & Johnson, 2008) or when they experience conflict between the presented concept model and their existing mental model. However, the expertise reversal effect did not hold true in their second experiment. In their second experiment, Craig et al. (2012) compared learning gains across the Q + E condition, the Q condition, and the monologue condition for both high level students (high school honors students) to students in a standard physics class. They found that the students in the Q + E condition learned significantly more from pretest to posttest than
students in the M condition and Q condition. Overall, the honors students learned significantly more than the standard students. Unlike the previous experiment, honors students learned most in the same Q + E condition, despite the additional content compared to the Q and monologue conditions.

**Trialogues in ITS**

AutoTutor has traditionally used a dialogue conversation framework, that is, one human learner and one animated pedagogical agent. More recently additional animated pedagogical agents have been included in AutoTutor interactions (Cai et al., 2011; Forsyth et al., 2013; Graesser et al., 2016; Graesser, Forsyth, et al., 2017; Graesser et al., 2018). Trialogues can be used in two different ways in AutoTutor. First, the added agent can act as a peer agent to the human learner. The peer agents can mirror a student’s current knowledge state by either agreeing with the student or by providing a response to a question that is similar to the student’s response. This allows for all negative feedback to be directed towards the peer agent and may avoid any undesirable effects negative feedback might have on the human learner (e.g., demotivation). Additionally, the peer agent can act as a model for the human learner by asking deep level reasoning questions and providing thorough self-explanations. The benefits of asking deep level reasoning questions and self-explanations can be obtained vicariously (Craig et al., 2000; Craig et al., 2006; Gholson & Craig, 2006). Researchers have assumed trialogue environments, both interactive and vicarious, are more appropriate for low-domain knowledge students. There is some support that trialogue-based ITSs are useful for promoting vicarious learning for low-level students, but the data is underwhelming.

While the focus of the proposed experiment is vicarious versus interactive dialogues in AutoTutor, the existing research on trialogues in ITSs provides insight into the effect of
vicarious modeling of student agents, as well as whether the implementation of a second peer or
student agent provides additional benefits over dialogues in ITSs. The content for the proposed
experiment is drawn from Operation ARA an AutoTutor based systems which is described
below. Adjustments to the Operation ARA system were required to fit the proposed design, but
the content has been professionally developed and evaluated in previous research.

Operation ARIES

ARIES (Acquiring Research Investigative and Evaluative Skills), also known as Operation
ARA (Acquiring Research Acumen), is a trialogue-based ITS. ARIES is an AutoTutor-based ITS
that teaches scientific reasoning in a serious game setting (Forsyth et al., 2013; Halpern et al.,
2012; Millis et al., 2011). ARIES utilizes several different teaching components including
AutoTutor dialogues, teachable agent trialogues, and vicarious observations of AutoTutor agents.
The “vicarious condition” is not purely vicarious in that the human students are occasionally
asked if they think the student agent is correct or incorrect. The “dialogues” are trialogues,
strictly speaking, in that a student agent does occasionally contribute to the conversation. In
ARIES, students work through chapters that cover scientific reasoning topics like “theories and
hypotheses”, “validity”, and “experimenter bias.” Before students interact with AutoTutor, they
work through an interactive e-text on a specific topic. They are then assessed within game and
are placed into one of the three AutoTutor conditions describe above, based on their
performance. Students who demonstrate a low-level understanding of the topic are placed into
the “vicarious-learning trialogue” condition. Students with intermediate knowledge are placed
into the “standard trialogue” condition, where they are tutored by the tutor agent while the peer
agent occasionally contributes to the conversation. Students who show high knowledge are
placed into the “teachable agent trialogue” condition. Overall, students in ARIES showed
significant learning gains from pretest to posttest and significantly outperform students who do
not use ARIES (Halpern et al., 2012). However, Halpern et al. (2012) did not observe any significant differences between the three conversation types. This could be interpreted (with caution) that students were appropriately placed within each conversation setting and so no one conversation framework was strictly better than the others. It is worth highlighting that students in the study were not randomly assigned to each conversation framework and instead were assigned to the conditions based on their levels of knowledge or understanding. Therefore, it cannot be reasonably determined which conversation framework is best for students with different levels of knowledge.

**AutoTutor CSAL**

More recently, AutoTutor trialogues have been used in *AutoTutor CSAL* (Center for the Study of Adult Literacy; Graesser et al., 2016; Shi et al., 2018). CSAL teaches reading comprehension strategies to adults with low literacy. Specifically, CSAL students learn how to comprehend text at multiple levels. For example, CSAL students are taught how to predict the purpose and structure of the text given different text signals, acquire vocabulary using contextual clues, and to summarize texts. Unlike other AutoTutor applications, students in CSAL do not type their responses to questions. Instead, the adult students select items on a screen in response to questions. For example, the students may be asked to highlight a specific sentence that helps indicate the topic of the text. In other situations, students respond by answering multiple choice questions. Each item is carefully constructed and logically map onto key knowledge components and misconceptions. The peer agent in CSAL sometimes agrees versus disagrees with the student responses. As in Operation ARIES, any negative feedback is redirected to the peer agent in order to prevent demotivation (Graesser, Forsyth, et al., 2017). Human students primarily receive neutral feedback (“OK, let’s think about this”) or positive feedback (“Great Job!”). In some
instances, the user is actively competing with the peer agent in a game. In this setting, whenever a human learner or peer agent answers a question correctly, they receive a point which is displayed on the interface. Here, the peer agent answers are adaptively selected to guarantee the human learner always wins or ties with the peer agent. Game-like features have been found to improve learning and motivation in other trialogue frameworks (Jackson & McNamara, 2013; Millis et al., 2017) and highlights the many roles the added peer agent can play. For example, in addition to a co-learner and a competitor, the second agent can also operate as someone who contradicts the other agent. This framework has been used to elicit confusion in human students. It turns out that college students learn more in conditions that generate confusion (Lehman et al., 2013). When students are confused, they are in a state of cognitive disequilibrium and often seek to restore equilibrium by reflecting carefully on the content or problem.
Chapter 4. Research Questions & Methodology

Research Questions

Currently, there are competing hypotheses that suggest that interactive tutoring should, in general, be more effective than vicarious learning (ICAP model; M. T. H. Chi et al., 2017, 2018; M. T. H. Chi & Wylie, 2014). However, there is evidence that certain vicarious conditions in ITSs can outperform interacting with the ITS (Craig et al., 2006). For human tutoring and other learning environments, there is strong evidence that indicates student prior-knowledge interacts with the type of instruction they receive (Aleven, Mclaughlin, et al., 2016; Cronbach & Snow, 1977; Kalyuga, 2014; Kozma, 1991). However, the results are mixed for learning in ITSs. ITS designers consider many different multimedia learning strategies and tutoring strategies. They apply dozens of these strategies at all levels of the system (e.g., student model, domain model, tutor model, and user interface). Generally, this strategy has proven effective. As ITSs become more sophisticated there are greater opportunities to apply entirely new modes of interaction (e.g., virtual-reality, augmented-reality, and simultaneous collaborative learning) and ways to detect and adapt to finer-grained learner features. However, it is still unclear how even simple design features interact with each other, as well as aptitude-treatment interactions. Carefully controlled experiments are needed to disentangle ATIs in ITSs.

The current experiment tests the following questions. Is student learning affected by the interaction between the type of tutoring framework and learner domain knowledge? Specifically, which tutoring framework, interactive dialogue or yoked-vicarious dialogue, is most effective for high versus low domain knowledge students? Do students perform best when observing tutees who share their knowledge level versus deviate from their knowledge level?
Conditions

Participants in the interactive dialogue condition (Interactive) interacted with the tutor agent, as described in the “Tutoring dialogue in AutoTutor” section. Participants in the yoked-vicarious condition saw a replay of a random participant in the interactive condition. Specifically, they were shown a recording of all on-screen interactions between a participant in the interactive condition and AutoTutor. Thus, the appearance of the interface remained the same as the interactive condition, but the student was not able to respond or interact with the tutor agent.
Hypotheses

The study design allowed the following hypotheses to be tested, as summarized in Table 1.

Table 1

Hypotheses with Participant Domain Knowledge and References

<table>
<thead>
<tr>
<th>Participant Domain Knowledge</th>
<th>Hypotheses</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>ICAP: interactive &gt; vicarious on posttest</td>
<td>Chi &amp; Wylie. (2014)</td>
</tr>
<tr>
<td>Low Prior Knowledge</td>
<td>Low prior knowledge participants &gt; high prior knowledge participants in vicarious condition on posttest</td>
<td>Craig et al. (2012)</td>
</tr>
<tr>
<td>Low Prior Knowledge</td>
<td>Low prior knowledge participants in vicarious condition &gt; low prior knowledge interactive on posttest</td>
<td>Craig et al., 2006; Driscoll et al., 2003; Graesser et al., 2017; Halpern et al., 2012; Millis et al., 2011</td>
</tr>
<tr>
<td>Low Prior Knowledge</td>
<td>In the vicarious condition, low prior knowledge participants who observe low prior knowledge participants who observe high prior knowledge participants on posttest scores</td>
<td>Kalyuga et al., 2001; Sweller &amp; Cooper, 1985</td>
</tr>
<tr>
<td>High Prior Knowledge</td>
<td>Interactive high prior knowledge participants &gt; vicarious high prior knowledge participants on posttest</td>
<td>Kalyuga, 2007a, 2014; Reisslein et al., 2006; Schnotz, 2010; Sweller et al., 2003</td>
</tr>
<tr>
<td>High Prior Knowledge</td>
<td>Vicarious high prior knowledge participants who observe high prior knowledge participants who observe low prior knowledge participants on posttest</td>
<td>Kalyuga, 2007a, 2014; Reisslein et al., 2006; Schnotz, 2010; Sweller et al., 2003</td>
</tr>
</tbody>
</table>

(1) Regardless of learner domain knowledge, if the ICAP framework holds true (M. T. H. Chi & Wylie, 2014), all students should learn better in the interactive dialogue condition than in the yoked-vicarious conditions. Three alternative predictions address low domain knowledge
participants and their learning in the vicarious conditions. (2) Low prior knowledge participants are expected to perform better than high prior knowledge participants in the vicarious condition. This pattern would align with the results from Craig et al. (2012) which found that low prior knowledge students learned more in the “questions + explanations” vicarious condition than high prior knowledge students in the same condition. (3) Low knowledge students are expected to perform better in the vicarious condition than in the interactive conditions. This hypothesis runs counter to ICAP which would predict low knowledge students in the interactive condition performing better than the vicarious conditions. This prediction is based on research that saw low knowledge students learn more from vicarious conditions than interactive conditions (Craig et al., 2006) and that vicarious learning is particularly helpful for low domain knowledge students (Driscoll et al., 2003). Additionally, developers of trialogue systems have tacitly assumed vicarious trialogue environments are best for low domain knowledge students (Graesser et al., 2017; Halpern et al., 2012; Millis et al., 2011). (4) Low knowledge students who vicariously observe the recordings of low knowledge students are expected to outperform those who observe replays of high knowledge students. The low knowledge student replays should provide a similar level of scaffolding needed by low knowledge student observers. That is, the increased amount of feedback observed in the low knowledge student replays may help resolve misconceptions of low knowledge student observers that would otherwise go unaddressed when observing the high knowledge student replays. Previous research has found that low knowledge students (or “novice learners”) require more specified and more frequent feedback (Kalyuga et al., 2001; Sweller & Cooper, 1985), which would be provided in the vicarious condition with low prior knowledge student replays and in the interactive condition.
Hypotheses 5a and 5b are concerned with high prior knowledge participants and the expertise reversal effect. (5a) High knowledge students are expected to perform best in the interactive condition, which is aligned with the ICAP hypothesis. (5b) High knowledge students who vicariously observe another high knowledge student tutoring interacting with AutoTutor are expected to outperform high knowledge students who observe low knowledge students interacting with AutoTutor. Per the expertise reversal hypothesis (Kalyuga, 2007a), the additional information presented in the low knowledge student videos (e.g., more feedback, hints, prompts) may increase the student’s cognitive load with redundant information as they attempt to work in the additional information into their existing mental model.

Materials, Participants, and Procedure

Content

Operation ARA is an ITS that teaches scientific reasoning concepts. The original Operation ARA was designed to be a serious game with an ongoing narrative about aliens who sabotage research with common research flaws, enhanced with triologues. The content of Operation ARA was broken down into three modules: “Basic Training”, where students worked through an e-textbook, “Proving Ground”, where students read a description of a study and were asked to identify flaws, and “Going Active” where students distinguish between flawed and unflawed research studies by asking questions (Millis, Graesser, & Halpern, 2014). Content from the “Proving Ground” module was adapted for use in the current study. Specifically, all story elements were removed from the content as well as the peer agent. The remaining content was delivered through a one-on-one AutoTutor dialogue environment that asks learners to read an example of a study and then to point out any flaws they find.

There were seven AutoTutor modules that cover the following ten concepts: independent and dependent variables, accuracy/precision/reliability, objective measuring, subject bias,
experimenter bias, variability/sample size, poor sample selection, premature generalization of results, correlation does not mean causation, and validity. Each module covered one to four different concepts. For example, the “Sugar Potatoes” module covered a study that investigated the role of sugar in maintaining the health of plants. The “Sugar Potatoes” study had four flaws: the dependent variable was not precise enough, the way the researchers measured the dependent variable was not objective, there was a poor sample selection, and the sample they did select was too small.

In AutoTutor, each flaw is designated as an “expectation” (as described in Section 3.3.1), each expectation has one to two hints, and each hint has one prompt. If the student’s response to the main question does not fully cover an expectation, the student receives a hint. If the student’s response to the hint was incorrect, they receive a prompt, which typically requires a one-word response. AutoTutor will then provide the correct answer and a brief explanation of the expectation. The process repeats if there is an additional expectation that was not covered by the original response. If all expectations are met, then AutoTutor provides a summary of the correct answer to the main question. All case studies are provided in the Appendix.

**Interface**

The AutoTutor interface used in the current study can be seen in Figure 3. The tutor agent is located at the top right of the screen. The main question is in the blue box at the top center of the screen. The case study is displayed throughout the interactions and is found in the center of the screen. Below the case study is a green box which displays the hint or prompt. Below the green
Participants

Participants were recruited through Amazon Mechanical Turk (MTurk), an online crowdsourced survey tool that allows “requesters” to post tasks or experiments and pay “workers” who complete the tasks. Participants received $14.50 for completing the experiment. MTurk allows requesters to screen potential workers based on several criteria. For the purposes of this study, MTurk workers were screened to include those who had completed 100 or more “tasks” with a 95% approval rating. This is a standard practice for behavioral research conducted on MTurk (Chandler et al., 2019; Peer et al., 2017). MTurk workers for this experiment were restricted to the United States and Canada. Participants were also recruited from The University of Memphis’
A total of 165 participants were recruited for the experiment. Of the 165 participants, 96 were included for analysis. Of the 69 participants that were not included in the analyses, 18 were interactive participants who “gamed the system” by providing random responses, one-word responses to each question, or generally responded in a way that indicated they were not attending to the task. Participants who spent less than 7 minutes on the posttest, 28 total, were not included for analysis. The 7-minute cut-off was selected because it takes an average reader seven minutes to read the posttest content, not including the amount of time needed to think of an answer to the question and enter the answer in the online quiz tool (Brysbaert, 2019; Carver, 1992). Participants who did not complete the experiment within the 3-hour time limit, 23 total, were not included for analysis. For the remaining 96 participants, 68 participants were recruited from MTurk and 28 were recruited from the subject pool. The average age of participants was 33.0. MTurk ($M = 37.5$) participants were older than SONA participants ($M = 21.9$) on average. For gender, 40 identified as female, 53 identified as male, two participants preferred not to answer, and one participant identified as non-binary.

**Assessment Tools**

All questions were multiple-choice questions with four options. The assessments used in the study included a total of 20 shallow level definitional questions and a total of 20 applied deep level reasoning questions. There were two shallow level questions and two deep level questions for each knowledge component covered in the didactic training and AutoTutor interactions. The counterbalanced pretests consisted of 10 of the 20 shallow level questions. The counterbalanced post-training tests consisted of all 20 shallow level questions. The posttest consisted of all 20
deep level questions, which briefly described a study with a flaw. Table 2 presents examples of both types of questions.

Table 2

*Key Concepts and Examples of Definitional and Applied Questions*

<table>
<thead>
<tr>
<th>Concept</th>
<th>Question Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Independent Variables, Dependent Variables, &amp; Hypotheses</strong></td>
<td>Definitional</td>
<td>In an experiment, the ____________ is being manipulated, changed, or controlled and the ____________ is used to measure an outcome or effect.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a. Subject Variable, Construct Variable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. <strong>Independent Variable, Dependent Variable</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. Construct Variable, Dependent Variable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>d. Hypothesis, Sample Variable</td>
</tr>
<tr>
<td><strong>Dependent Variables: Reliability, Accuracy, Precision</strong></td>
<td>Applied</td>
<td>Does training improve the ability to lift weights for young men? To answer this question, a researcher had a random sample of young men train in the gym by lifting heavy weights for one month. She then measured their ability to lift a two-pound ball. Is this a good measure?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a. Yes, because it immediately followed the training.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. No, because the study is unrelated to the study’s central concept.</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>c. No because the measure is not sensitive enough for this study.</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>d. Yes, because the men may not have been lifting the same amount of weight during training</td>
</tr>
<tr>
<td><strong>Subject Bias</strong></td>
<td>Definitional</td>
<td>Sometimes participants in a study alter their behavior because they know the purpose of the study. This phenomenon is referred to as</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a. poor sample selection</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>b. subject bias</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. attrition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>d. experimenter bias</td>
</tr>
<tr>
<td><strong>Subject Bias</strong></td>
<td>Applied</td>
<td>Ms. Snipplenose was hired by a drug testing firm to see if taking the cold remedy Cold-B-Gone really helped people get over their colds faster than they would have without the remedy. She used a sample of 100 people all of whom were just starting to get a cold. She randomly assigned them to either take the remedy or do nothing. She found that those who took the remedy reported that they felt better after 6 days and those who did nothing felt better after 7 days. What is one flaw in this study?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a. <strong>Subject bias.</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. Small sample size.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. Too large of a sample size.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>d. Experimenter bias.</td>
</tr>
</tbody>
</table>

*Note.* Each definitional question is used either once (pretest only) or twice (both pretest and post-training), depending on test order. Applied questions appear once in the post-test.
A demographics survey was administered to determine participant age, gender, education level, whether they are a native English speaker, the number of content related courses they had taken in the past, college or high school GPA, and ACT score.

**Interaction Replays**

All data regarding interactions with AutoTutor were stored in a Learning Record Store (LRS). Learner responses to questions, feedback, hints, pumps, and prompts were all recorded as xAPI statements. This includes the response time for each interaction, with millisecond accuracy. xAPI statements enable various learning systems to capture data in a standardized format (specifically with expressions specifying nouns, verbs, and objects of actions) which are stored in the LRS. Figure 4 shows an example of what a typical xAPI statement looks like when stored in the LRS.

![Figure 4](image.png)

*Figure 4. An excerpt of an xAPI statement as recorded in the LRS.*

Here, the noun is the “actor”, which in this case is the participant, SPR21PNT1227. The “verb” is “Response”, and the “object” is an AutoTutor activity. The question is a hint provided by AutoTutor in the Liver module. The “Response” is the participant’s response to the hint. The “latency” refers to the amount of time (34s) it took for the participant to hear the hint, think about an answer, and type the answer into the input box. These statements were retrieved from
the LRS and automatically applied directly within AutoTutor as input, and functionally created a “replay” for a vicarious participant to observe.

Figure 5 is a screenshot of what a vicarious participant would see when viewing the replay.

![Figure 5](image.png)

**Figure 5** A screenshot of an AutoTutor “replay” viewed by a vicarious participant.

In these replays, vicarious participants also saw the case study presented in the center of the screen, the hints that were provided to the original interactive participant were also displayed in a green box at the lower portion of the screen, and the main question was also located at the top of the screen in the blue box. Additionally, they saw the input the interactive participant provided being “typed out” as if the interactive participant was interacting with AutoTutor in real time.
Procedure

The experiment was conducted online using the survey tool, Qualtrics, in addition to AutoTutor. Participants first received an informed consent followed by the counterbalanced 10-item multiple choice pretest. Participants were then assigned into the interactive dialogue condition or a yoked-vicarious condition. Participants in both groups then worked their way through a set of slides on the concepts that appear in the AutoTutor lessons. Afterwards, participants completed a counterbalanced 20-item multiple choice post-training test. Participants were given 20 minutes to complete the post-training test. Participants were then assigned to either the vicarious condition or interactive condition. A form of block-randomization was necessary to build a surplus of interactive participant replays for vicarious participants to observe. Specifically, the first 10 participants were all assigned to the interactive condition so that the next block of 10 participants could be randomly assigned to either the interactive condition or the vicarious condition. Not all interactive participants could be used as interactive replays for various reasons (e.g., providing the same response for each question, response times short enough to indicate they were not attending to the material, or providing random letters as input). This made it possible to run out of replays for future vicarious participants, so it was necessary to assign participants to the interactive condition when enough viable replays were available to continue randomly assigning participants into the two conditions.

Participants assigned to the interactive dialogue condition worked their way through 7 interactive modules with AutoTutor. Each module displayed a case study or article describing a scientific experiment with an average length of 380 words. Participants in the vicarious condition saw a “replay” of an interactive participant’s interactions with AutoTutor. Each vicarious participant was randomly assigned to view a different interactive participant’s replay, except for
one instance where two vicarious participants viewed the same interactive participant’s replay due to a technical error. After participants in both groups finished viewing or interacting with AutoTutor, they were given 30 minutes to complete the 20-item multiple choice posttest with deep questions. Finally, participants were debriefed and provided a completion code to confirm they finished the study.
Chapter 5. Results

T- Sample Differences on Prior Knowledge and Assessment Times

*MTurk & SONA*

One concern with collecting data from different sources is that there could be differences between the two samples that affect learning outcomes that are not controlled. A series of independent samples t-tests were conducted to determine whether pretest, post-training, and posttest scores were significantly different between the MTurk and SONA groups. The t-tests yielded no significant differences, suggesting that there were no sample differences on the assessments. Table 3 shows the results of the t-tests comparing the MTurk group to the SONA group on pretest, post-training, and posttest scores.

Table 3

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>MTurk</td>
<td>68</td>
<td>.595</td>
<td>.218</td>
<td>-1.29</td>
<td>.203</td>
</tr>
<tr>
<td></td>
<td>SONA</td>
<td>28</td>
<td>.647</td>
<td>.164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-Training</td>
<td>MTurk</td>
<td>68</td>
<td>.732</td>
<td>.223</td>
<td>0.30</td>
<td>.766</td>
</tr>
<tr>
<td></td>
<td>SONA</td>
<td>28</td>
<td>.718</td>
<td>.180</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posttest</td>
<td>MTurk</td>
<td>68</td>
<td>.576</td>
<td>.206</td>
<td>0.26</td>
<td>.244</td>
</tr>
<tr>
<td></td>
<td>SONA</td>
<td>28</td>
<td>.524</td>
<td>.189</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The number of relevant courses taken is relatively similar between the MTurk and SONA participants, with 27% of MTurk participants having taken one or more domain-relevant courses, and 39% of SONA participants having taken one or more domain-relevant courses. The higher percentage of SONA participants who have completed one or more domain-relevant course is
expected, given that some participants were psychology majors who were currently taking or have already taken a required domain-relevant course. However, it appears that having taken a relevant course provided no advantage over those SONA participants who have not taken a relevant course. An independent samples t-test was conducted to determine if there were any differences between the SONA participants who had completed one or more domain-relevant course and those that have not completed a domain-relevant course on pretest scores. The results indicate that participants who have taken a relevant course \((M = .570, SD = .150)\) had significantly lower pre-test scores than those who had not taken a relevant course \((M = .697, SD = .156)\), \(t(22.15) = -2.16, p = .042\). Interestingly, the same trend was observed for MTurk participants. An independent samples t-test was conducted to determine if there were any differences between MTurk participants who completed a domain-relevant course and participants who had not taken a domain-relevant course on pretest scores. The results indicate that MTurk participants who had completed one or more domain-relevant course \((M = .458, SD = .234)\) had significantly lower pretest scores than participants who had not completed a domain-relevant course \((M = .648, SD = .188)\), \(t(26.5) = -3.16, p = .004\). The same trend was observed for all participants.

Another concern with collecting data from both SONA and MTurk is that the two groups were compensated differently, with MTurk participants receiving monetary compensation and SONA participants receiving SONA credit. These different incentives may have led to motivational or effort differences. MTurk workers’ primary reason for working on MTurk is, unsurprisingly, monetary compensation (Litman et al., 2015). MTurk workers may try to strike a balance between the amount of time they spend on a task and the amount of effort they put into the task, so their work gets approved. A 2011 study found that participants could be “over
compensated”, in that the high compensation rates might attract individuals looking to game the
system to quickly receive payment (Bohannon, 2011). However, more recent research shows that
data quality is not directly related to compensation for US-based participants (Litman et al.,
2015).

A previous study compared MTurk to a university participant pool on “instances of
insufficient effort responding” (IER) and found that the university students had slightly higher
rates of IER on two of IER indices but had similar IER rates on three of the five indices (Toich et
al., 2021). Participants who rushed through the three assessments could be considered as
“gaming the system” and attempting to receive compensation with minimal effort. For the
present study, an independent samples t-test was conducted to determine whether the MTurk
sample was different from the SONA sample on total time spent on assessments. The results
indicate that MTurk participants ($M = 30$ min $41$ s, $SD = 13$ min $18$ s) did not spend significantly
less time on their assessments than the SONA participants ($M = 25$ min $12$ s, $SD = 12$ min $36$ s),
t$(52.94) = 1.91$, $p = .62$). Taken together, these results suggest that the MTurk and SONA
samples do not appear to differ in a way that may have influenced the results of the following
analyses.

**Vicarious versus Interactive Conditions**

An independent samples t-test was conducted to assess whether there was a significant
difference on pretest scores between the vicarious and interactive conditions. The t-test yielded
no differences between the vicarious group ($M = .592$, $SD = .223$) and the interactive group ($M$
$= .629$, $SD = .183$) on pretest scores $t(91.85) = -0.88$, $p = .38$. Similarly, there was no significant
difference between the vicarious condition ($M = .696$, $SD = .251$) and the interactive condition
($M = .761$, $SD = .152$) on post-training scores $t(79.64) = -1.55$, $p = .125$. 
The vicarious condition had a higher percentage of participants who completed one or more
domain-relevant classes (41%) than the interactive condition (21%). An independent samples t-
test was conducted to determine whether this difference was associated with higher pre-test
scores for vicarious participants. Vicarious participants who had taken a relevant course \( (M = .470, SD = .214) \) performed significantly worse than vicarious participants who had not taken a
relevant course, \( (M = .676, SD = .190) \), \( t(37.64) = -3.46, p = .001 \). However, no significant
difference was observed between interactive participants who had taken a previous course \( (M = .557, SD = .204) \), and interactive participants who had not taken a previous course \( (M = .648, SD = .174) \), \( t(12.77) = -1.30, p = .218 \). While the difference is not significant, the mean
difference follows the same trend as the previous analyses on the number of classes and pretest
scores.

An independent samples t-test was conducted to determine if the vicarious group was
different from the interactive group on total time spent on assessments. The results indicate that
interactive participants \( (M = 27 \text{ min 12s}, SD = 10 \text{ min 35s}) \) did not differ significantly from
vicarious participants \( (M = 30 \text{ min 12s}, SD = 15 \text{ min 27s}) \) on the total time spent on the
assessments \( t(85.21) = 0.84, p = 0.405 \). These results suggest that before the vicarious and
interactive participants entered the AutoTutor conditions, they did not appear to differ in a way
that may have influenced the results on posttest scores.

**Group Differences on Learning**

*Prior Knowledge on Post-Training, Posttest Scores*

Each participant was labeled as having high versus low prior knowledge based on a median
split on pretest scores. To assess the potential differences on posttest performance between prior
knowledge groups (high vs low), an independent samples t-test was conducted. The high prior knowledge group ($M = .672, SD = .171$) significantly outperformed the low prior knowledge group ($M = .485, SD = .186$) on posttest scores, $t(86.09) = -5.08, p < .001$. High prior knowledge participants ($M = .840, SD = .101$) also outperformed low prior knowledge students ($M = .651, SD = .231$) on the post-training test, $t(82.10) = -5.49, p < .001$.

**Effects of Knowledge on Learning Behaviors in AutoTutor**

It was predicted that high prior knowledge participants would interact with AutoTutor differently than low prior knowledge students. Specifically, high prior knowledge participants would receive less feedback, and therefore give fewer but more correct responses than low prior knowledge participants.

Pretest scores do not appear to be a significant predictor of the number of responses provided in AutoTutor, $F(1,45) = 2.42, p = .127$, adjusted $R^2 = .03$. However, it may be more informative to know if the participant’s knowledge state (as determined by post-training scores) going into the AutoTutor conditions predict the number of responses provided when interacting with AutoTutor. A linear regression was conducted to determine the effect of a participant’s knowledge state going into the interactive AutoTutor condition on the number of responses given in AutoTutor. The results indicate that post-training scores are a significant predictor of the number of responses provided in AutoTutor, $F(1, 45) = 12.21, p < .001$, with an $R^2$ of .213. This suggests that as post-training scores increase, participants provided fewer responses ($\beta = -0.46, p = .001$), and therefore received less feedback, and were more accurate with their responses than participants with lower prior knowledge heading into AutoTutor. Figure 6 depicts a scatter plot of post-training scores on the number of responses provided by the participant in AutoTutor.
The number of responses provided to AutoTutor in the interactive condition also significantly predicted posttest scores, $F(1, 45) = 10.48, p = .002$, with an $R^2$ of .189. This suggests that a participant’s higher performance in AutoTutor predicts their performance on the posttest ($\beta = -0.44, p = .002$). Taken together, these results highlight that prior knowledge does affect behaviors in AutoTutor, which are then reflected in the interactive replays that vicarious participants observe. Also, performance in AutoTutor predicts performance on posttest. Figure 7 depicts a scatter plot of the number of responses provided by the participant to AutoTutor on posttest scores.
Condition Differences on Posttest Scores

An independent samples t-test was conducted to determine whether the interactive group outperformed the vicarious group on posttest scores. The results indicate that the interactive group ($M = .601, SD = .178$) significantly outperformed the vicarious group ($M = .522, SD = .217$) on the posttest $t(91.686) = -1.96, p = .0265$. However, when controlling for pretest scores in an analysis of covariance, the results were not significant. Table 4 shows the means and standard deviations of the two conditions for pretest scores, post-training scores, and posttest scores.

Table 4

Means and Standard Deviations for Conditions on Pretest and Posttest Scores

<table>
<thead>
<tr>
<th>Condition</th>
<th>Pretest</th>
<th>Post-Training</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Vicarious</td>
<td>.592</td>
<td>.223</td>
<td>.696</td>
</tr>
<tr>
<td>Interactive</td>
<td>.628</td>
<td>.183</td>
<td>.761</td>
</tr>
</tbody>
</table>
A one-way ANCOVA was conducted to assess the effect of learning environment (interactive vs vicarious) on posttest scores, while controlling for pretest scores. Prior to analysis, it was determined that there was a linear relationship between pretest and posttest for both groups, as assessed by visual inspection of a scatter plot. There was homogeneity of regression slopes because the interaction term was not statistically significant, $F(1,92) = .75, p = .390$. Standardized residuals for the two groups were normally distributed, as assessed by Shapiro-Wilk’s test ($p > .05$). There was homoscedasticity, assessed by visual inspection of the standardized residuals plotted against predicted values. After adjustment for pretest scores, no significant difference was observed between the two conditions on posttest scores $F(1, 93) = 3.05, p = .084, \eta^2 = .032$, although the difference would be significant according to a one-tailed statistical test.

**Condition Differences with Prior Knowledge on Posttest Scores**

A one-way ANCOVA was run to assess the effect of group type on posttest scores when also considering prior knowledge, as determined by pretest scores. Specifically, this analysis would help differentiate posttest performance between knowledge groups in both the interactive and vicarious conditions (i.e., high prior knowledge in vicarious, low prior knowledge in vicarious, high prior knowledge in interactive, low prior knowledge in interactive) while controlling for pretest scores. Prior to the analysis, it was determined that there was a linear relationship between pretest and posttest for all groups, as assessed by visual inspection of a scatter plot. There was homogeneity of regression slopes because the interaction term was not statistically significant, $F(3,88) = .55, p = .648$. Standardized residuals for the two groups were normally distributed, as assessed by Shapiro-Wilk’s test ($p > .05$). There was homoscedasticity, assessed by visual inspection of the standardized residuals plotted against predicted values. After
adjustment for pretest scores, no significant difference was observed between the four groups on posttest scores, $F(3,91) = 1.59, p = .197, \eta^2 = .050$.

An additional one-way ANCOVA was run to assess the effect of group type on posttest scores when considering prior knowledge but controlling for post-training scores instead of pretest scores. The median split on pretest scores was used to determine prior knowledge of the four groups. Prior to analysis, it was determined that there was a linear relationship between pretest and posttest for all groups, as assessed by visual inspection of a scatter plot. There was homogeneity of regression slopes because the interaction term was not statistically significant, $F(3,88) = .16, p = .690$. Standardized residuals for the two groups were normally distributed, as assessed by Shapiro-Wilk’s test ($p > .05$). There was homoscedasticity, assessed by visual inspection of the standardized residuals plotted against predicted values. There was homoscedasticity, assessed by visual inspection of the standardized residuals plotted against predicted values. After adjustment for post-training scores, a significant difference was detected between the four groups on posttest scores $F(3,91) = 3.53, p = .018, \eta^2 = .104$. Table 5 shows the estimated marginal means of the four domain knowledge groups with post-training as a covariate. The results show that the high prior knowledge interactive participants significantly outperformed low prior knowledge vicarious participants on posttest ($p = .015$). All other comparisons were not significant.
Table 5

*Estimated Marginal Means of Posttest Scores as a Function of Pretest Subgroups with Post-Training as Covariate*

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Vicarious</td>
<td>.485</td>
<td>.033</td>
<td>[.421, .549]</td>
</tr>
<tr>
<td>High Vicarious</td>
<td>.607</td>
<td>.038</td>
<td>[.532, .682]</td>
</tr>
<tr>
<td>Low Interactive</td>
<td>.547</td>
<td>.031</td>
<td>[.486, .607]</td>
</tr>
<tr>
<td>High Interactive</td>
<td>.649</td>
<td>.038</td>
<td>[.572, .724]</td>
</tr>
</tbody>
</table>

The previous analysis used pretest scores to determine prior knowledge for the 4 groups. The following ANCOVA uses post-training tests to determine prior knowledge for the 4 groups with pretest scores as a covariate. After adjusting for pretest scores, in contrast to the previous analysis, both high and low prior knowledge interactive participants significantly outperformed the low prior knowledge vicarious group \((p = .002\) and \(p = .005\), respectively), \(F(3,91) = 3.95\) \(p = .011\), \(\eta^2 = .115\). Additionally, the high prior knowledge vicarious group outperformed the low prior knowledge vicarious group on posttest scores \((p = .004)\). Table 6 presents the means, standard errors, and confidence intervals of the four subgroups.

Table 6

*Estimated Marginal Means of Posttest Scores as a Function of Post-Training Subgroups with Pretest as Covariate*

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Vicarious</td>
<td>.435</td>
<td>.040</td>
<td>[.355, .515]</td>
</tr>
<tr>
<td>High Vicarious</td>
<td>.594</td>
<td>.031</td>
<td>[.531, .656]</td>
</tr>
<tr>
<td>Low Interactive</td>
<td>.580</td>
<td>.033</td>
<td>[.515, .646]</td>
</tr>
<tr>
<td>High Interactive</td>
<td>.608</td>
<td>.034</td>
<td>[.540, .676]</td>
</tr>
</tbody>
</table>
**Vicarious Pair Groups**

Participants in the vicarious condition were broken down into four subgroups based on the vicarious participants prior knowledge, as determined by pretest scores, and their “target’s” prior knowledge (i.e., the prior knowledge of the observed interactive participant). This was conducted to assess whether the prior knowledge of the observed interactive learner affects the learning of the vicarious learner, and to assess whether some knowledge pairs learn better than others, and how these vicarious groups compare to interactive participants. For example, by breaking the vicarious learners into these subgroups, we can assess whether low prior knowledge vicarious participants learned more from high prior knowledge interactive participants than they did with low prior knowledge interactive participants. An ANCOVA was conducted to assess whether there were differences between the vicarious knowledge subgroups as well as interactive participants on posttest scores while controlling for pretest scores. There was homogeneity of regression slopes as the interaction term was not statistically significant, $F(4, 86) = 1.35, p = .257$. Standardized residuals for the group types were normally distributed, as assessed by Shapiro-Wilk’s test ($p > .05$). There was homoscedasticity, assessed by visual inspection of the standardized residuals plotted against predicted values. After adjustment for pretest scores, no significant difference was observed between the five groups on posttest scores $F(4,90) = 1.32, p = .267, \eta^2=.056$. Since the differences between the adjusted means were not significant, a Bonferroni post hoc evaluation of simple effects was not conducted. Table 7 shows the means and standard deviations of each subgroup.
Table 7

*Estimated Marginal Means of Posttest Scores as a Function of Pretest Subgroups with Pretest as Covariate*

<table>
<thead>
<tr>
<th>Group Type</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% C.I</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive</td>
<td>.593</td>
<td>.025</td>
<td>[.543, .643]</td>
<td>47</td>
</tr>
<tr>
<td>Low Vicarious Low Interactive</td>
<td>.497</td>
<td>.044</td>
<td>[.410, .583]</td>
<td>18</td>
</tr>
<tr>
<td>Low Vicarious High Interactive</td>
<td>.503</td>
<td>.052</td>
<td>[.400, .606]</td>
<td>11</td>
</tr>
<tr>
<td>High Vicarious Low Interactive</td>
<td>.600</td>
<td>.054</td>
<td>[.493, .708]</td>
<td>11</td>
</tr>
<tr>
<td>High Vicarious High Interactive</td>
<td>.543</td>
<td>.059</td>
<td>[.427, .660]</td>
<td>9</td>
</tr>
</tbody>
</table>

It may be more informative to consider the participant’s updated knowledge state immediately before entering the AutoTutor conditions than it is to consider their prior knowledge before beginning the experiment. The didactic training may have had a greater impact on participants’ learning behaviors in AutoTutor than participants’ prior knowledge entering the study. It should be noted that participants who had taken domain relevant courses before beginning the experiment did not outperform those who had not taken a domain relevant course. Instead, the opposite was true for all participants; those who had not taken a previous domain relevant course outperformed those who had taken a domain relevant course on pretest scores. Participants’ knowledge directly coming into the study may not be as informative as their knowledge leading into the AutoTutor conditions. If prior knowledge is an important factor for predicting learning behaviors in both the interactive and vicarious AutoTutor conditions, then the participants’ current knowledge state (as reflected in the post-training scores) should not be ignored.

Participants in the vicarious condition were broken into four different subgroups based on their prior knowledge, as determined by a median split on post-training scores, and their
“target’s” prior knowledge. The interaction group was also included in the analysis, making five groups total. There was homogeneity of regression slopes as the interaction term was not statistically significant, $F(4, 86) = 1.98, \ p = .104$. Standardized residuals for the group types were normally distributed, as assessed by Shapiro-Wilk’s test ($p > .05$). There was homoscedasticity, assessed by visual inspection of the standardized residuals plotted against predicted values. After adjustment for post-training scores, there was a statistically significant difference in posttest scores between the 5 groups, $F(4,90) = 4.26, \ p = .003, \ \eta^2 = .159$. Post hoc analysis was performed with a Bonferroni adjustment. The interactive group significantly outperformed the Low Vic. High Int. subgroup (i.e., low prior knowledge vicarious participants who watched high prior knowledge interactive participants) on the posttest (mean difference of .208, $p = .004$). Participants in the High Vic. Low Int. subgroup also had significantly higher posttest scores than participants in the Low Vic. High Int. subgroup (mean difference of .254, $p = .004$). All other simple effects were non-significant. Table 8 presents the estimated marginal means, standard errors, and confidence intervals of each subgroup.

Table 8

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive</td>
<td>.593</td>
<td>.023</td>
<td>[.548, .639]</td>
</tr>
<tr>
<td>Low Vicarious Low Interactive</td>
<td>.497</td>
<td>.055</td>
<td>[.388, .607]</td>
</tr>
<tr>
<td>Low Vicarious High Interactive</td>
<td>.385</td>
<td>.051</td>
<td>[.283, .486]</td>
</tr>
<tr>
<td>High Vicarious Low Interactive</td>
<td>.639</td>
<td>.043</td>
<td>[.553, .725]</td>
</tr>
<tr>
<td>High Vicarious High Interactive</td>
<td>.546</td>
<td>.041</td>
<td>[.464, .628]</td>
</tr>
</tbody>
</table>
An additional one-way ANCOVA was conducted, like the previous ANCOVA (using post-training scores to determine prior knowledge), but with the interactive group split into two new subgroups (High Prior Knowledge Interactive and Low Prior Knowledge Interactive). Prior to analysis it was determined that there was homogeneity of regression slopes as the interaction term was not statistically significant, $F(5,84) = 1.13, p = .352$. Standardized residuals for the group types were normally distributed, as assessed by Shapiro-Wilk’s test ($p > .05$). There was homoscedasticity, assessed by visual inspection of the standardized residuals plotted against predicted values. After adjustment for post-training scores, there was a statistically significant difference between the six groups, $F(5, 89) = 3.44, p = .007, \eta^2 = .162$. Post hoc analysis was performed with a Bonferroni adjustment. As with the previous ANCOVA, the Low Vic. High Int. group had the lowest average posttest scores, which were significantly lower than three subgroups. The High Vic. Low Int. subgroup outperformed the Low Vic. High Int. subgroup on posttest scores (mean difference of .257 (95% CI, .048, to .466) $p = .005$). Low prior knowledge interactive participants also significantly outperformed the Low Vic. High Int. group on posttest scores (mean difference of .198 (95% CI, .17 to .38), $p = .021$. Finally, the high prior knowledge interactive participants outperformed the Low Vic. High Int. subgroup on posttest scores (mean difference of .223 (95% CI, .034 to .413), $p = .009$. These results replicate the previously described ANCOVA, but also show that both low prior knowledge interactive participants and high prior knowledge interactive participants outperformed the Low Vic. High Int. group on posttest scores.

_Vicarious Domain Knowledge Distances Effect_

The analysis of covariance on categorical prior knowledge pairs (as determined by pretest scores) yielded null results but it may still be informative to examine the continuous delta, or
prior knowledge “distance”, between the vicarious participant and their interactive pair. Prior knowledge distance was calculated by subtracting the interactive participant’s pretest score from their paired vicarious participant’s pretest score (i.e., Vicarious Pretest – Interactive Pretest). Negative values indicate that the observed interactive participant’s pretest score was higher than the observer’s pretest score. A linear regression was conducted to determine the effect of the knowledge distance on pretest performance between a vicarious participant and their observed interactive pair on posttest performance. The results indicate that knowledge distance on pretest scores significantly predicted posttest performance, $F(1,47) = 17.84, p < .001$, accounting for 27.5% of the variance in posttest scores with an adjusted $R^2 = .260$. Figure 8 is a scatter plot showing the effect of pretest distance on posttest scores, with blue points depicting observed high prior knowledge participants and gray points depicting the observed low prior knowledge participants.
Another linear regression was conducted using post-training scores instead of pretest scores to calculate prior knowledge distance’s effect on posttest performance. Distance was calculated by subtracting the interactive participant’s post-training score from their paired vicarious participant’s post-training score (i.e., Vicarious Post-training – Interactive Post-training). Negative values indicate that the observed interactive participant’s post-training score was higher than the observer’s post-training score. Knowledge distance on post-training scores significantly predicted posttest performance, $F(1,47) = 24.80, p < .001$, accounting for 34.5% of the variation in posttest scores for vicarious participants with an adjusted $R^2 = .331$. Figure 9 is a scatter plot showing the effect of post-training distance on posttest scores.
Figure 9. Scatter Plot of Post-Training Distance on Posttest Scores.
Chapter 6. Discussion

The goal of this study was to determine how prior knowledge may interact with various learning environments in a conversation-based ITS. Another goal was to assess whether the prior knowledge of an observed tutee may affect the learning of the observer, and whether this effect changes given the prior knowledge of the observer.

The first hypotheses predicted that ICAP framework may be replicated. The ICAP framework suggests that, in general, interactive learning environments should be more effective than constructive learning environments, which should be more effective than active learning environments, all of which should be more effective than passive environments (M. T. H. Chi & Wylie, 2014). Replicating ICAP would predict that the interactive condition would consistently outperform the vicarious condition, regardless of the participant’s domain knowledge, or the domain knowledge of the observed tutee. When using post-training scores to determine prior knowledge, the low prior knowledge interactive group significantly outperformed the low prior knowledge vicarious group. A similar trend was observed for high prior knowledge interactive participants (estimated marginal mean = .608) and high prior knowledge vicarious participants (estimated marginal mean = .594), although the difference was not significant. Taken together, these findings support the ICAP framework.

Further analysis provided a more nuanced understanding of the above results. When breaking the vicarious groups into four new subgroups, the low prior knowledge interactive group significantly outperformed low prior knowledge participants in the vicarious condition that observed high prior knowledge participants. There was no significant difference between the low prior knowledge interactive group and low prior knowledge participants who observed other low prior knowledge participants. However, while not significantly significant, low prior knowledge
participants who observed other low prior knowledge participants (estimated marginal mean = .494) performed worse on the posttest compared to the low prior knowledge interactive group (estimated marginal mean = .581), which is a trend that is consistent with ICAP. Further analysis indicated that the difference between high prior knowledge interactive participants and high prior knowledgevicarious participants, while statistically insignificant, was greater when comparing high prior knowledge interactive participants (estimated marginal mean = .607) to high prior knowledge vicarious participants who saw other high prior knowledge participants (estimated marginal mean = .547), which is a trend that does support ICAP. Interestingly, in the same analysis an opposite trend was observed, which ran counter to the ICAP hypothesis. Specifically, while not statistically significant, the trend shows that high prior interactive participants had lower average posttest scores than high prior knowledge vicarious participants who observed low prior knowledge vicarious participants (estimated marginal mean = .640). ICAP was replicated when comparing high prior knowledge interactive participants to high prior knowledge vicarious participants who observed other high prior knowledge participants.

The observed trends largely support ICAP, except for the case where high prior knowledge vicarious participants observed low prior knowledge participants interact with AutoTutor. One explanation for this exception could be high prior knowledge participants benefitted from the increased presence of conflict episodes in the condition where they saw low prior knowledge participants, which outweighed the potential interference of the redundant information they observed. Whereas high prior knowledge interactive participants did not experience as many of these potentially beneficial conflict episodes when interacting with AutoTutor.

The second hypothesis predicted that low prior knowledge participants would perform better than high prior knowledge participants in the vicarious condition. This would have replicated the
findings of Craig et al. (2012). In their study, the “Q + E” condition involved participants watching a tutorial dialogue with deep level reasoning questions and with a tutee providing explanations to further expand on tutor statements. Low prior knowledge participants outperformed high prior knowledge students in the same condition. Vicarious participants in the current study saw a tutor ask a question and a tutee provide responses, and the tutor provide feedback to the tutee responses. The results indicated that, when dividing students into domain knowledge groups by post-training scores and controlling for pretest scores, there was a significant difference between low prior knowledge participants in the vicarious condition ($M = .435$) and high prior knowledge participants in the same condition ($M = .594$) on posttest scores, but in favor of the high prior knowledge vicarious group.

The third hypothesis predicted that low knowledge students would learn more from the vicarious condition than they would from the interactive condition. While this hypothesis runs counter to ICAP, it was supported by research that shows vicarious learning environments in tutoring are helpful for learners (Craig et al., 2000; Driscoll et al., 2003; Gholson et al., 2009; Graesser, Cai, et al., 2017; Millis et al., 2017; Muldner et al., 2014) and are sometimes more effective than interactive conditions (Craig et al., 2006). Also, developers of ITSs that use a triologue format have assumed low prior knowledge participants benefit from the presence of a tutee agent they can observe and model (Graesser, Forsyth, et al., 2017; Millis et al., 2011). The results of this study show that the opposite may be true. It appears low prior knowledge participants, as determined by post-training scores, in the interactive condition ($M = .580$) outperformed those in the vicarious condition ($M = .435$). The significantly greater posttest scores of low prior knowledge interactive participants compared with low prior knowledge vicarious participants replicates the results of Craig et al. (2004). In both of their experiments,
Craig et al. (2004) did not include high prior knowledge participants. In this case, their yoked-vicarious condition is essentially the same as the low vicarious watching low interactive subgroup (LL) in the current study. They also found a significant difference between their yoked-vicarious and interactive conditions on learning gains, in favor of the interactive condition. It is important to note, however, that the significantly greater posttest scores of the low prior knowledge interactive group in the current study appears to only apply to the low prior knowledge vicarious subgroup that observed high prior knowledge participants (LH). There was no significant difference between low prior knowledge interactive participants and the LL group. The finding that there was no significant difference between low prior knowledge interactive participants and the LL subgroup replicates the results of the Craig et al. (2006) experiments, which also only included low prior knowledge participants. Craig et al. (2006) did not observe a significant difference between their interactive and yoked-vicarious, essentially the LL group, in both experiments. These results did not confirm the third hypothesis, but they support the findings of Craig et al. (2006), while also providing some support for the ICAP framework.

While these results do not confirm the third hypothesis, trialogues may still provide benefits over dialogues for low knowledge students because they often incorporate both interactive and vicarious learning. The combination of these two modes may reduce the negatives of a purely vicarious environment (e.g., increased opportunities to interact with the system may prevent learner disengagement). The current study had no built-in attention checks, beyond providing an end-of-lesson password to move forward, so there was no guarantee that the vicarious participants were attending to the modules. Trialogue learning environments that incorporate both opportunities for vicarious learning and interactive learning may have a distinct advantage over vicarious learnings alone.
The fourth hypothesis predicted that low prior knowledge students in the vicarious condition would benefit more from observing other low prior knowledge students interact with AutoTutor than watching high prior knowledge students. Low prior knowledge participants should receive more scaffolding through hints, pumps, and prompts than their high prior knowledge counterparts because they would be less accurate when interacting with AutoTutor. First, it was important to establish that low prior knowledge students do indeed perform differently in AutoTutor than high prior knowledge students. Results of a regression showed that prior knowledge (as measured by post-training scores) was a significant predictor of the number of responses given when interacting with AutoTutor. The number of responses increased as post-training scores decreased. This indicates that low-prior knowledge participants in the interactive condition were less accurate when responding to AutoTutor, and therefore received more feedback, hints, and prompts than high prior knowledge participants. It is also worth noting that fewer responses in AutoTutor predicted posttest outcomes, which shows that behaviors in AutoTutor are important indicators of both the participant’s knowledge state and their learning. The post hoc comparisons of an ANCOVA revealed that, while the difference was not significant, low prior knowledge participants who watched other low prior knowledge participants did have higher average posttest scores than low prior knowledge participants who viewed high prior knowledge participants. This trend does appear to support the fourth hypothesis but warrants further investigation.

Both hypotheses 5a and 5b were grounded in the expertise reversal hypothesis. The expertise reversal hypothesis suggests that high prior knowledge students’ learning can be disrupted when they are exposed to redundant information (Kalyuga, 2014). This redundant information can come in the form of frequent feedback, guidance, and instruction. As mentioned earlier,
participants who entered the AutoTutor interactive condition with lower prior knowledge had low accuracy and received more feedback, hints, pumps, and prompts than participants with higher prior knowledge. The increased number of interactions with AutoTutor may have disrupted the learning of high prior knowledge participants.

Hypothesis 5a predicted that high prior knowledge participants would perform best in the interactive condition, because higher accuracy in AutoTutor results in less redundant hints and feedback than they may observe when watching other participants interact with AutoTutor. This would also align with ICAP. Interactive learning in AutoTutor has a better chance of keeping high prior knowledge learners in their zone of proximal development because they receive feedback and further questions only on material for which they need support.

The results of an ANCOVA on the four prior knowledge subgroups (Low Vicarious, High Vicarious, Low Interactive, High Interactive) using post-training scores as a covariate, indicated that although high prior knowledge interactive participants had higher mean posttest scores ($M = .649$) than high prior knowledge vicarious participants ($M = .607$), the difference was not significant. This suggests that high prior knowledge participants may not have been disrupted by the increased feedback they saw in the vicarious condition. This could be in part due to the exposure of more “conflict episodes” (Chi et al., 2017) when observing other participants interact with AutoTutor. Specifically, high prior knowledge participants may have had misconceptions that were addressed when observing the mistakes made from other high prior knowledge participants.

Hypothesis 5b predicted that vicarious high prior knowledge participants who viewed other high prior knowledge participants would outperform high prior knowledge participants who viewed low prior knowledge participants. Low prior knowledge participants were less accurate
in the interactive AutoTutor condition, which means they received more feedback and hints than high prior knowledge participants. Therefore, the expertise reversal hypothesis would suggest that high prior knowledge participants who watched other high prior knowledge participants interact with AutoTutor would outperform high prior knowledge participants who viewed low prior knowledge participants. Interestingly, while statistically insignificant, it was found that high prior knowledge participants who watched high prior knowledge participants ($M = .546$) had lower average posttest scores than high prior knowledge participants who watched low prior knowledge participants ($M = .639$). High prior knowledge participants did not appear to be disrupted by redundant information presented in the vicarious condition and may have even benefited from the added hints, feedback, and conflict episodes in AutoTutor. However, further research is needed to differentiate the effects of additional AutoTutor feedback and hints on learning in vicarious settings.

Most of our analyses split participants into subgroups based on their prior knowledge. However, a regression analysis was also conducted to place the knowledge differences between the vicarious pairs on a continuous scale. The knowledge distance between the observer and the participant they observed was calculated by subtracting interactive participants’ pretest scores from the vicarious participants’ pretest scores. Negative values indicated that the vicarious participant’s knowledge was lower than the participant they observed interact with AutoTutor. The results indicated that as distance increased in a positive direction (e.g., the vicarious participant’s prior knowledge was higher than the participant they observed), the observed participant’s prior knowledge played a smaller role in the performance of the observer. As the vicarious participant’s knowledge decreased and the prior knowledge of the observed tutee increased, the posttest scores decreased for the observer. The same trend was observed when
looking at prior knowledge based on post-training scores. The results of an ANCOVA showed similar results: low prior knowledge participants who viewed high prior knowledge participants performed significantly worse than all other subgroups. These results also lend further support to the idea that high prior knowledge participants in vicarious settings benefit more from observing more hints and feedback, with potentially redundant information and conflict episodes, than they do from observing a peer with fewer hints and feedback.

**Implications**

There was some support for the ICAP hypothesis, which predicted that learning in the interactive condition would be more beneficial than learning in the vicarious condition, regardless of the learner’s prior knowledge. However, the current study found more support that the effectiveness of learning environment depends partly on the learner’s prior knowledge, and in the case of vicarious learning, the prior knowledge of the observed tutee. The results of the current study suggest that in some situations, vicarious learning can be as effective as interactive learning. For example, there was no significant difference on posttest scores between high prior knowledge participants in the interactive condition and high prior knowledge participants in both vicarious condition subgroups (high prior knowledge watching high prior knowledge, high prior knowledge watching low prior knowledge). However, while not significant, the high prior knowledge interactive group, which was predicted to be the best learning environment for high prior knowledge participants, had a lower average posttest performance than high prior knowledge participants who viewed low prior knowledge participants, which was predicted to be the worst learning environment for high prior knowledge participants. It appears that the increased amount of potentially redundant information that was observed by high prior knowledge participants who were paired with low prior knowledge participants may have
improved learning outcomes instead of disrupting learning outcomes, as seen in the expertise reversal hypothesis literature.

ITS developers that wish to provide effective learning environments for high prior knowledge participants should not discount the potential benefit of vicarious environments. When considering how to develop effective triologues (i.e., one human learner, one virtual tutee, one virtual tutor) for high prior knowledge learners, the current study provides some support for the tutee agent to represent a low prior knowledge learner. The potentially redundant information and conflict episodes that a low prior knowledge tutee agent provides may still be beneficial to high prior knowledge participants.

This study’s results suggest that the worst learning environment for low prior knowledge participants is a vicarious setting in which they observe high prior knowledge participants in AutoTutor. This is likely because a participant with lower prior knowledge requires support in the form of hints and corrective feedback, which they would see less of when watching a high prior knowledge participant interact with AutoTutor. It appears that observing accurate and positive learning behaviors in AutoTutor is not enough to promote learning for low prior knowledge vicarious participants. This could be explained by the zone of proximal development (Chalkin, 2003; Vygotsky, 1987). Low prior knowledge participants who viewed high prior knowledge participants were likely observing answers to questions outside of their ZPD. Specifically, low prior knowledge participants required more assistance and guidance from AutoTutor, which was not available when observing high prior knowledge participants. For the content to remain in their ZPD, they would need to observe other low prior knowledge participants interact with AutoTutor, because they would observe AutoTutor provide more guidance and assistance to the low prior knowledge interactive participant. Further, the results
indicate that as the knowledge gap between low prior knowledge vicarious participants and their observed high prior knowledge participant increased, posttest scores for the vicarious participant decreased. ITS developers interested in providing effective triologue environments for low prior knowledge learners should consider keeping the tutee agent’s knowledge at or below the human learner’s level. Observing good answers alone does not appear to promote learning for low prior knowledge learners.

The finding that prior knowledge can predict the number of responses the learner will provide in AutoTutor, and that performance in AutoTutor predicts performance on posttest scores has important implications for adaptive conversation-based ITSs. First, a relatively simple metric, the number of times a participant responds to AutoTutor, can be a powerful tool for understanding their current knowledge state. This is an intuitive finding, given that dialogue-based ITSs typically attempt to keep students in their zone of proximal development. Tutor agents move on to new questions and material when a student provides a correct response. This also functions as quick “working as intended” check for conversation-based ITSs. Second, this study found that the fewer number of responses provided in AutoTutor predicts outcomes on a posttest that contained all transfer problems. The posttest items assessed the participant’s ability to apply their understanding of research methods and concepts of scientific inquiry to new scenarios in the form of different case studies. Less frequent responses in AutoTutor appears to be a significant predictor of deeper learning.

Limitations

The posttest was useful for assessing the participants’ ability to transfer their knowledge from the AutoTutor interventions to new problems, but it prevented direct assessment of learning gains from pretest and post-training to posttest. The time to complete the study ranged from 1.5
to 2 hours, so an effort was made to keep the assessments at reasonable lengths to prevent test fatigue. However, an alternative assessment strategy would be to include new shallow level items in the posttest as well as the deeper level transfer questions. This would allow for pretest to posttest assessment of learning gains by separating the posttest shallow items from the deeper level transfer items for analysis.

There are several potential issues with using MTurk as a recruitment tool for behavioral research. The quality of MTurk response data has been shown to be in decline since 2018 (Chmielewski & Kucker, 2020). However, there is evidence that MTurk data quality between MTurk and a sample of college undergraduates were similar, with the college students showing higher instances of “insufficient effort responding” instances on two indices of IER qualities (Toich et al., 2021). There is some evidence that indicates that as the length of the task increases, the number of good workers decreases and the number of bad workers increases (Gadiraju et al., 2015). Higher paying tasks may be attractive to MTurk workers looking to game the system, but research shows that higher MTurk pay is associated with better performance on tasks (Aguinis et al., 2021; Casey et al., 2017; Rogstadius et al., 2021). The payment for this study was set by the U.S. minimum wage per hour for workers, as recommended by Aguinis et al. (2021). No significant differences were observed on total assessment time taken and assessment scores between the MTurk and SONA sample. However, there may still be motivational differences between the two samples given the different types of compensation used for both groups, with MTurk workers receiving monetary payment and SONA students receiving SONA credit.

MTurk tracks the approval rating of requesters which is visible to MTurk workers. The average approval rating of the current study on MTurk steadily declined over time. As workers saw a lower approval rating, the sample may have changed as well. Specifically, workers who
were more risk-averse may have avoided signing up for the task because they saw a lower approval rating. Alternatively, there may have been an increase in the number of workers who were less risk-averse and were willing to risk a drop in their task completion approval rating. Future work with MTurk would benefit from tracking detailed information regarding their approval rating and how this may change the behavior of their MTurk workers.

An effort was made to create “replays” that were exact copies of the interactive participant’s behavior. This comes with the advantage of providing a potentially more ecologically valid learning environment for the vicarious participants, but it also has some drawbacks. For example, participants in the vicarious condition may have read more slowly than their interactive counterparts. This may have resulted in participants in the vicarious condition trying to catch up to the learning session by skimming the reading materials, potentially ignoring feedback provided to the interactive participant. Future studies using a similar methodology should consider including an additional condition for vicarious participants, where the amount of time available to read the case studies was determined by the vicarious participant instead of the interactive participant.

Attention checks were not included in the interactive condition because it was easy to track user activity within AutoTutor. Attention checks were not used for the vicarious condition because they may have distracted the observer from viewing the interactive replay and created an arbitrary difference between the two conditions on the flow of learning. However, this did result in the potential for vicarious participants to disengage during the replays. The argument can be made that vicarious conditions are inherently less engaging than interactive learning conditions and inserting attention checks during learning creates a learning environment that is not entirely vicarious.
Finally, another limitation of the study regards the sample size for subgroups. For example, the high prior knowledge vicarious participants who observed other high prior knowledge students had an $n$ of 9. The vicarious participants were randomly assigned to their interactive participant, so there was no control in place to balance the sample sizes for each subgroup.

**Future Directions**

The learning of high prior knowledge participants did not appear to diminish by an increase in redundant information. This runs counter to the expertise reversal hypothesis and research regarding redundancy and cognitive load (Kalyuga, 2007a, 2007b). It raises the question as to whether there is an optimal amount of redundancy in the form of corrective feedback and hints for high prior knowledge vicarious participants. More specifically, at what point does redundant information for high prior knowledge vicarious learners interrupt their learning, if at all? Further studies exploring vicarious learning in the context of prior knowledge should consider manipulating the amount of feedback presented to high prior knowledge vicarious learners to determine when, if at all, the expertise reversal effect applies.

Low prior knowledge vicarious participants who viewed high prior participants interact with AutoTutor were the lowest performing subgroup. This could have been caused by a lack of hints and feedback necessary to establish the scaffolding low prior knowledge participants needed. Vicarious participants can model positive learning behaviors presented by virtual tutees (Craig et al., 2000; Gholson, et al., 2009), but the current results suggest that observing good responses to AutoTutor questions is not enough to promote learning for low prior knowledge participants. There may be an optimal ratio between the amount of feedback observed and the correctness of the feedback for low prior knowledge vicarious learners. The more accurate a person is in AutoTutor, the less hints and feedback they receive. However, the AutoTutor rules can be
adjusted so that all hints and prompts are presented to the learner even if they are providing all correct responses. This may result in a frustrating learning environment for high prior knowledge participants, but low prior knowledge vicarious participants could benefit from observing both correct answers and an increased number of hints and prompts. It may be the case that the learning benefits of observing conflict episodes is simply due to the increased amount of learning material that follows an incorrect statement. If this is the case, low prior knowledge participants should perform as well or better in an “all content, all correct” vicarious condition than they would in an “all content, all incorrect” vicarious condition.

The method of creating “replays” of interactive learners in AutoTutor by pulling data from an LRS and using it as input in a separate instance of AutoTutor creates the potential for individualized vicarious learning environments. The method used in this dissertation can be expanded to leverage real student interaction data in a way to automatically generate vicarious tutoring conversations for observation, given both the target student’s aptitude and knowledge state. By expanding on this approach, conversations between a tutor agent and tutee agent can be assembled from various students’ responses to generate the “ideal” tutee agent for any given observer. Correct or incorrect statements can be drawn from the pool of correct or incorrect statements in the LRS and then presented to the observer as responses to AutoTutor questions in both dialogue and trialogue environments. When answers are evaluated by AutoTutor, they need to meet a specific threshold of semantic overlap (as determined by LSA). It may be the case that vicarious learners can learn best by observing partially correct answers, or answers that are near misses (e.g., answers that approach the threshold without meeting it). Responses that are entirely irrelevant can be discarded from the pool of past interactive responses in the LRS. A finer-
grained understanding of the impact of tutee behaviors on vicarious learning can be gained by manipulating the “correctness” of the observed responses.

**Conclusion**

The current study provides support for the idea that prior knowledge and learning environment (i.e., interactive dialogues or vicarious dialogues) affects learning outcomes. Particularly, low prior knowledge learners are better off interacting with AutoTutor than they are with observing high prior knowledge learners. In fact, the worst condition for low prior knowledge learners was observing high prior knowledge learners. Regarding previous research, there were some conflicting results between the Craig et al. (2004) experiments and the Craig et al. (2006) experiments. Specifically, the Craig et al. (2004) experiments saw low prior knowledge participants performing significantly better in an interactive condition than in a yoked-vicarious condition where they observed other low prior knowledge participants. This study replicated the results of Craig et al. (2006), which saw no significant difference on learning between their low prior knowledge participants in the interactive condition and low prior knowledge participants who viewed other low prior knowledge participants in their yoked-vicarious condition. For high prior knowledge participants, the expertise reversal hypotheses predicted that they would perform worse in conditions where there was an increased amount of feedback, hints, and conflict episodes. However, the current study did not find support for this hypothesis. While the difference was not significant, the trend shows that high prior knowledge participants were not disrupted by this redundant information, feedback, and conflict episodes, but rather may have even benefited from them. Further research is needed to determine the effect of potentially redundant information, an increase in conflict episodes, and an increased amount of feedback on high prior knowledge learning in intelligent tutoring systems.
References


among eighth to eleventh graders in the domains of computer literacy and Newtonian physics.


https://doi.org/10.1016/j.chb.2017.01.038


https://doi.org/10.1111/apps.12344


Appendix

All case studies included in the experiment borrowed from the Operation ARA project.

Sugar & Potatoes

Global warming may have taken yet another toll upon nature: millions of flowers around the world have begun to die. Researchers and scientists are rushing to find out the exact cause and possible ways to counteract this catastrophe. “The main concern is that with a loss of flowers, there will be a great decrease in oxygen production, which in turn will be quite a problem for humanity,” stated Dr. Ryan Kimball, one of the many scientists researching the loss of flowers.

Many scientists believe that this loss is partially caused by a decrease in turgor pressure inside the plants. Turgor pressure in plant cells helps flowers stay strong, rigid, and upright. But, researchers disagree on how to increase turgor pressure. Some say it’s the ratio of sugar present in the surrounding environment to sugar present in the plant that affects turgor pressure, but that is uncertain.

In a simple study, scientists have recently tested whether increases in sugar present in the environment decrease the turgor pressure of potatoes, a common plant. “If turgor pressure turns out to be responsible for this problem, we can assume that more intense rays from the sun due to global warming have increased the amount of sugar depleted from the plants relative to the environment, thus causing the issue at hand,” explained senior researcher Aubrey Trast.

For the study, four test tubes were filled with sugar water, one with 0% sugar, another with 1% sugar, a third with 3% sugar, and the final tube filled with 7% sugar. A scientist carefully selected and cored 4 healthy potatoes and put each of the single potato cores into each tube. After a week delay, the results came in.

Scientists recorded the rigidity of the potatoes using a well-known Turgor-Scale in which they bend the potato core and judge its flexibility. They found the following turgor scores for the tubes with 0%, 1%, 3%, and 7% sugar: 3.7, 3.7, 3.6 and 3.5 centimeters for each test tube. The researchers reported that these slight differences were not significantly different from each other.

“We believe that given these results, we can tentatively conclude that sugar does not affect turgor pressure in potatoes, but further research must be done to support this point,” explained Trast. “And if similar findings are made, we will know that we cannot grow stronger flowers by simply removing sugar from the soil or water. We will need to look into another explanation for diminished turgor pressure.”

Modified Food

Recently, researchers have explored whether genetically modified food might actually taste better than natural food. Dr. Longert, an expert researcher working for a company that genetically modifies food, conducted an experiment to test whether consumers prefer the taste of genetically modified food over unaltered food.

Longert decided to conduct the experiment himself to ensure it was conducted according to his methodology. The experimenter selected 50 individuals who worked at the company to participate in the experiment. “Dr. Longert approached me over lunch break a few months ago,” says colleague Sandra Winert, “and I was really excited about his study and was happy to help out. In fact, all of us who participated were excited about the importance of this study for our company!”

The 50 coworkers were randomly assigned to either a genetically modified food group or the natural food group. This resulted in 25 participants assigned to each group, one containing those coworkers that were consuming genetically modified foods on a regular basis (GM group) and one that consumed natural foods (N group).
For the actual experiment, all fifty participants were asked to look at a variety of food items that were either genetically modified (GM group) or natural (N group). These foods included produce, such as bell peppers and apples, as well as processed foods made of either genetically modified or natural foods. All 50 participants rated their food items on a 5 point scale, indicating how visually appealing the food looked (with 1 indicating “not visually appealing at all” to 5 indicating “highly visually appealing”).

The results indicated clearly that the participants in the GM group rated their foods, all of which were genetically modified, as significantly more appealing than participants in the N group. Dr. Longert decided to replicate the study with a new sample of coworkers, and found similar results. “The data have spoken clearly,” says Dr. Longert. “We have to assume at this point that genetically modified food probably tastes better than natural food.”

Butterflies Are Not Free

JACKSONVILLE -- Butterflies are beautiful creatures who some believe personify the soul. In Chinese mythology, two butterflies flying together represent love. In Japan, a swarm of them signify a bad omen. In the United States, they represent big bucks. Wedding planners are buying butterflies by the thousands so that newly married couples can witness thousands filling the sky as they say their vows. Wealthy homeowners buy them to fill their gardens. Butterflies are becoming an industry to themselves.

But, there is a problem. That is, although butterflies grow on trees, many desirable species are in low supply. Here enters Jim Cowan, a high school biology teacher in Florida. Mr. Cowan used his biology class to test whether planting several sweet pepperbrush shrubs in a 20 mile area near his school would attract butterflies and keep them in the area.

“These shrubs were going to be planted by the park district anyway so I thought I would test whether they would increase the butterfly population in the area,” explained Cowan. “If they did, then the kids would be happy knowing that they would have butterflies in their near future and even at their weddings.”

At the beginning of the semester (1 month before the shrubs were planted) his class placed insect friendly traps throughout the area. At the beginning of the semester (1 month before the shrubs were planted) his class placed insect friendly traps throughout the area. They were careful not to touch the butterflies or mark them in any way so they would not disturb them. Weekly recordings continued after a whopping 10,000 shrubs were planted in the park, and they ended at the completion of the semester, about 2 months later.

“However, something occurred in the middle of the semester that did not ensure butterflies in the students’ future other than in their stomachs. We experienced a devastating drought coupled with high temperatures. It only lasted a week, but it adversely affected the health of the plants in the areas. Many of the pepperbrush shrubs died,” Cowan lamented.

Fortunately, the study did not die out like the shrubs. “It appears that the butterflies in fact must have been attracted to the shrubs because when the shrubs died out, so did the butterfly population. There was a huge statistically significant drop in the number of butterflies before and after the drought,” revealed Cowan.

Cowan is so excited about his discovery that pepperbrush shrubs attract butterflies that he plans to raise butterflies at his uncle’s farm and sell them to nearby wedding planners.

Facilitated Communication Helps Autistic Children

Marie Anderson, Staff Writer

Autistic children face severe social, communicative and behavioral impairments. Not surprisingly, researchers of autism have begun to cast a wide net for finding ways to improve the communication skills of people with autism.
One promising procedure is known as facilitated communication, in which a person gently holds or touches an autistic child’s hand while he/she is writing. It is believed that autistic children respond positively to physical contact resulting in improved communication.

The positive effects of facilitated communication were recently demonstrated in a study published in the journal *Autism*. Professor Art Jones, a leading autism researcher who ran the study, said that he was “interested in finding out if something as simple as physical contact with another human would help autistic children become better at something they have so much trouble with and something that is so essential to everyday life-communication with others.”

The study employed parents of autistic children to serve as experimental assistants. These volunteers had been previously trained in facilitated communication and believed in its therapeutic value. “I needed help to run the children through the study, and the volunteers were happy to do so,” said Jones.

Forty-four autistic children from Illinois participated in the study. These children were randomly assigned to either an experimental or a control group. In the experimental group, the volunteers gently held the children’s hands as they wrote. In the control group, the same volunteers were present and sat next to the children as they wrote, but they did not have any physical contact with them.

Following the experiment, researchers who were unfamiliar with the hypothesis analyzed what the children wrote. Specifically, they looked at the complexity of the children’s writing. Sentence complexity was defined on a number of variables, including the length of the sentences, and the number of words at or beyond the grade level of the child, both standard measures of writing complexity.

The results showed that the children in the experimental group (those exposed to facilitated communication) wrote significantly more complex sentences than those in the control group. For Jones, the results of this study are very encouraging, but he is planning to replicate the study later this year before promoting the program as a therapy.

---

Heavy Metal Music: A Teenager’s Perspective

Like a lot of other kids my age, I like to listen to heavy metal music. Like a lot of other parents, mine complain about the type of music I listen to all the time. They think that my listening to heavy metal music is going to make me sad and depressed. I always tell them they’re wrong-my friends and I all listen to heavy metal music, and we are some of the happiest people I know! I knew that the only way to get them to back off was to prove to them that they were wrong, so I did some research. I found the one and only article that seems to be out there about the effects of heavy metal music on mood.

In this study, 80 participants were randomly put into one of two groups. In one group, the participants were told that they could not listen to heavy metal music for one whole month. In the other group, the participants had to listen to heavy metal music for two hours every day for a month. After the month was over, everyone had to fill out a survey with a lot of different rating questions. There were so many different questions that no one who participated could have guessed that depression was being studied.

The main question that the researchers were interested in was how the participants rated their anxiety/emotional level. So the key question was “What is your current emotional state,” and the participants answered by circling either “anxious” or “not anxious.” The researchers were unaware of which condition the participants were in when they did the final data analyses. The researchers found that 60% of the participants who listened to heavy metal music reported feeling anxious, as did 58% of the participants who did not listen to the music at all. This was not a statistically significant difference.

So, what does all this mean for me? Well, it means that even though my parents admitted to me that I was right—I’m not going to become depressed from listening to heavy metal music-they still don’t want me to
listen to it. But, the take-home message for all the other teens and their parents out there is this: Listening to heavy metal music is not going to lead to depression, so listen to it as much as you like!

Liver Health Research

Most people these days know how important a healthy liver is for general health. The liver works around the clock to protect us by eliminating any toxic substances from our body. When individuals are exposed to too many toxins, for example by living in an environment that poses health risks, or consuming a high rate of unhealthy foods, the liver can become overworked and will not be able to deal with the toxins effectively. At that point, we become vulnerable to diseases.

Recent studies have observed an alarming increase in liver-related problems for the general population. However, to date, there has been no convenient and tasty food to support a healthy liver. Lemons are great, but few people enjoy consuming lemons.

Fortunately, liveRplus Industries appears to have succeeded in creating a delicious new health drink, Liv-Pro, which is intended to supply the body with the nutrients needed to boost liver health and encourage the elimination of toxins.

Larry Fretnot, executive director of liveRplus Industries, funded a series of studies to investigate the effectiveness of Liv-Pro. One-hundred-twenty participants were recruited for this study.

All participants were workers at liveRplus Industries and had easy access to the beverage in the cafeteria and from vending machines throughout the factory grounds at a discounted rate.

The scientists conducting the study distributed a questionnaire to the volunteers which assessed the amount of Liv-Pro consumed daily for two weeks. Based on these data, average consumption per person was calculated.

To test for liver health, researchers looked into the eyes of a person. This is because the clarity of the iris, the colored band around the eye’s pupil, is an indicator of liver health. Thus, to measure liver health, iridology was used, in which a judgment is made on the color of the eye’s iris and this is indicated on a clarity scale.

The results of the iridology tests in the current study showed a correlation between the variables of interest: liver health increased with higher consumption of Liv-Pro.

We are encouraged by these first results, Mr. Fretnot announced Friday. Variations in liver health was a result of Liv-Pro consumption! We believe our beverage will lead to healthy livers across the world, but first we are currently planning a study with the general population to follow up on this study.

Are There Real Benefits to Using Antibacterial Soap?

The purpose of all soaps is to clean our skin. But clearly, not all soaps are created equal. Shelves are packed with soaps in all colors and shapes, some containing lotion, some containing exfoliating agents; some scented, some hypoallergenic.

If you have felt overwhelmed by the choices, you are not alone. Most people have expressed uncertainty when shopping for soap. One question that has plagued consumers for some time now is whether it makes a difference to buy antibacterial soaps.

Antibacterial soaps seem to be the default choice of public offices and institutions in order to curb the spread of harmful bacteria. But, is this kind of soap really superior to regular soap in effectively getting rid of harmful germs? The answer appears to be no. Dr. Jelex and her team from a well-respected university recently obtained evidence refuting the higher effectiveness of antibacterial soap over plain soap.
The researchers set out to compare the number of colony forming units (CFUs; a standard bacterial count) in the palms of eight participants before and after they washed their hands with either antibacterial or plain soap, as determined by random assignment, for two minutes.

The soap was presented in no-name containers, thus the participants were not aware of the condition to which they were assigned. The number of CFUs was estimated based on samples taken by swab from a predetermined part of the palm for all participants. The swabs were labeled with numbers that allowed for blind scoring and later matching to the condition.

The CFU numbers of the four participants in the antibacterial soap group were reduced by 76; however, there was a reduction of 74 for the four participants in the plain soap group. The difference was not significant, according to Jelex.

On the basis of this finding, “we strongly recommend that consumers save their money and reach for the plain soap options,” advises Dr. Jelex.