THESIS

EXAMINING THE ROLE OF AUTOMATION TRANSPARENCY IN LEARNING WITH INTELLIGENT TUTORING SYSTEMS

Submitted by:

Rebecca L. Pharmer

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Master of Science

Fort Collins, Colorado

Summer 2023

Master's Committee:

Advisor: Benjamin Clegg

Christopher Wickens
Rosa Martey
Sara-Anne Tompkins

ABSTRACT

EXAMINING THE ROLE OF AUTOMATION TRANSPARENCY IN INTELLIGENT TUTORING SYSTEMS

In the present study, a training system that either assigned restudy of concepts based on learner performance (adaptive instruction) or provided a set amount of restudy (static instruction) was designed to investigate whether adding automation transparency into an intelligent tutoring system would improve learning outcomes in an assembly task. Participants received instruction on the assembly process of 8 unique shapes. They were provided with error sensitive feedback that served the transparency manipulation, where some participants received explanations of why they were receiving restudy or were given generic feedback. Findings indicate that adaptive instruction may be most beneficial to learning when automation transparency provides learners with an understanding of how the system is responding to their performance. Findings and implications to be discussed.

ACKNOWLEDGEMENTS

I would first like to acknowledge my advisor, Dr. Benjamin Clegg, and my committee, Dr. Chris Wickens, Dr. Rosa Martey, and Dr. Sara-Anne Tompkins. The completion of this thesis would not have been possible without all the encouraging and helpful feedback I received from each of them. Thank you all for your commitment to my success and for pushing me to realize my potential.

I would also like to acknowledge the support of my family and friends. Thank you to my lab mates, Amelia Warden and Colleen Patton, for always being available to help when I needed it. Thank you to my parents, Lorrie and Jim Pharmer, for always listening to my research ideas and being such a motivating support system from across the country. Lastly, I need to thank my husband, Braddock McDonald, for taking great care of our little family while I pursue my graduate education.

TABLE OF CONTENTS

INTRODUCTION

Computerized training has granted educators the ability to reach a larger student population than traditional face-to-face instruction. Advancements in educational technology have allowed for instructional systems to mimic one-on-one human tutors by providing individualized training to learners across a variety of domains. Remote delivery of instructional content was cornerstone during the 2020 Coronavirus pandemic, as seventy-seven percent of public schools reported moving classes to online distance-learning formats (U.S. Department of Education, National Center for Education Statistics, 2022). This recent uptick in the need for economical remote instruction with the propensity to accommodate learners at all levels calls for an increased understanding of the methods that will best support learning in virtual environments. The present study aims to address how empirical evidence from the human automation interaction literature can aid in improving automated educational technologies.

Intelligent Tutoring Systems (ITS)

Intelligent tutoring systems have been defined in the literature as a type of automated system that can adapt to the learner's aptitude to provide an optimal level of support and challenge (Durlach & Lesgold, 2012). This is accomplished through automated measurement of the learner behavior while using the system, an analysis of that behavior to create a model of the learner's competency in the domain, and then utilizing this model to determine how the learning content needs to adapt for optimal learning of the individual (Durlach, 2019). Tailored training reduces the cognitive burden of the learner to accurately assess their own performance and adjust study methods accordingly (Mayer, 2014; Sweller, 1991). Nwana (1990) describes ITS as the intersection between the disciplines of computer science, psychology, and education and

training. As technology advanced in the mid twentieth century, researchers at the intersection of these domains came to the realization that instructional content could be generated and scored by computers. Yazdani (1986) criticized early use of these systems, asserting that a computer cannot have the domain knowledge of a human and could not explain high level questions to learners. Today, ITS compose of sophisticated algorithms that model domain knowledge from teams of experts. Though building these systems a large undertaking, modern ITS can execute a variety of instructional strategies comparable to how a human tutor may change their training approaches with skilled versus novice learners.

There are a number of ways that an ITS can adapt training to the learner (see Durlach et al, 2011 for a review). For example, adjusting the difficulty of content to become easier or harder based on the learner's responses. This might include increasing the speed or number of targets in a particular trainer (Tseng, Chu, Hwang, & Tsai, 2008). Some systems operate by the adapting the spacing of content, giving less frequent presentation of content the user demonstrates knowledge of (Metzler-Baddeley & Baddeley, 2009; Mettler et al., 2016). Similarly, using "mastery criteria" or a set of requirements for proficiency allows learners to master-out or drop already mastered concepts from their study (Mettler et al., 2020; Whitmer et al., 2020). Another intervention would be the use of metacognitive prompts where the system instructs the learner to think about their current state and identify which areas need improvement, allowing for guided self-correction of errors (Schwonke et al., 2006; Pon-Barry et al., 2006). Further, providing error-sensitive feedback in the form of a score (Johnson et al., 2017) or description of learning objectives not yet mastered. Although effective, consideration must be taken to ensure the feedback is easily understandable, delivered at intervals that are not intrusive to the training task, and provide helpful instruction. The present study aims to focus on error-sensitive feedback, as it

allows the system to inform the user of not only their current learning performance, but also the reasoning behind the decision of the system to reassign presentations of content.

As described above, consideration for designing training systems with error-sensitive feedback is the type of feedback being presented. Feedback can be presented in a multitude of ways, conveying very general performance information to very detailed corrective information. This phenomenon is discussed in Kluger and DeNisis' (1996) review on the effects of feedback interventions during training, where the authors examined the trend of feedback interventions having a negative impact on performance when learners are provided with feedback that is more directed to their own abilities rather than being more task-focused. Building upon this review, Johnson & Priest (2014) established the Feedback Principle of Multimedia Learning, stating that novice participants learn better with explanatory feedback recommending a change in task strategy than feedback indicating performance alone. They define outcome feedback as informing learners of the accuracy of their response or performance, while process feedback is defined as proving the learner with an explanation as to why their response was accurate or inaccurate. In other words, process feedback aims to direct the learner on ways to change their strategies in order to improve performance while outcome feedback aims to make the learner aware of their performance. Support for the use of process feedback over outcome feedback is further demonstrated in a study by Astwood et al. (2008), where the authors investigated how each type of feedback influenced decision making in a call-for-fire simulation game. The game required participants to decide which targets to destroy first based on a set of prioritization criteria. They found that participants given process feedback on how to better form their decisions outperformed participants who received outcome feedback in the form of performance scores with no instruction on how to tailor their decision making or strategy. The researchers

attributed this finding to performance feedback better informing the learner about how they should be making decisions. Taken together, these findings indicate that error-sensitive feedback provides learners with some awareness of their own performance, which allows them to

Another consideration when designing a training system with error-sensitive feedback is how often to present the learner with feedback. While many studies have explored different feedback timing interventions, the results seem to be context specific which giving little generalizable advice for when to present feedback. This is possibly due to conflicting operational definitions of immediate versus delayed (Johnson et al., 2017). For example, a study by, Johnson et al. (2013) examining different feedback schedules during a serious game-based simulation found no significant differences in game scores for participants presented with immediate versus delayed feedback but reported that participants received marginally higher scores in the immediate feedback condition, but ultimately reported no significant difference. They also found that novice participants reported a higher cognitive load in the delayed feedback condition. While the findings of the self-report measure may not be conclusive alone, they do call for more research into whether delaying feedback increases extraneous cognitive load for novices learning procedures in a simulation-based task.

Automation transparency

ITS can be classified as an automated system within Parasuraman et al.'s (2000) Levels of Automation. They obtain information from the learner, then calculate and execute the most optimal direction for the user. A known characteristic of automation that has influence on users' compliance with automation is the level of transparency the system provides (Sargent et al, 2023). Transparency can be thought of as the amount of information a system provides about its decision-making process to the user of the system to allow the user to develop an accurate mental

model of the agent (Bhaskara et al., 2021). A recent meta-analysis by Sargent et al. (2023) found an a strong positive effect of transparency on performance in studies comparing transparent systems to non-transparent control systems. The amount and type of information needed to optimize performance may change based on the domain or context in which an automated agent is used, as these measures also have the capacity to overload the user, resulting in reduced effectiveness of transparency, creating a decrement in task performance or impacting users trust and compliance with the agent (Bhaskara et al., 2021; Sweller, 1991). Taken together, we can infer that transparency is overall useful for performance, the parameters in which it is implemented must fit the associated task. Parameters for transparency implementation have yet to be fine-tuned, however, some have developed frameworks for implementing transparency. For example, Chen's model of transparency (2014), that conceptualizes transparency according to Endsley's (1995) levels of situational awareness: perception of elements in the environment; comprehension of those elements; and projection of their status in the near future. The SAT model provides guidance for implementing transparency into a system to promote better human-autonomy teaming by increasing the situational awareness of the operator. The model consists of three components:

1. The purpose of the agent: What is it trying to achieve?
2. The process of the agent: Why is it recommending an action?
3. The performance of the agent: What will happen if the recommendation is followed?

While this model is usually applied to automated systems meant to support the operator in a task where both the agent and operator share common goals, it has yet to be applied to automated agents in the education realm. In the context of an Intelligent Tutoring System, this could be

adapted to provide transparency of the ITS to the learner. For example, if we apply Chen's (2014) model to the learning domain, we will provide the learner with:

1. Purpose of the agent: The system's goal is to maximize learner retention of content.
2. Process of the Agent: The system is providing a recommendation for restudy based on learner performance during training and comparing it to optimal performance at test.
3. Performance of the agent: If the recommendation is followed, the user can expect greater learning outcomes. Disregarding the recommendation may lead to forgetting content.

Research related to the role automation transparency in educational technologies seems to be a relatively unexplored domain, aside from a few recent studies. Putnam and Conati (2019) conducted a limited sample pilot study to examine user attitudes toward transparency in an ITS and gain insight into the types of explanations users want out of a training tool. In this study, participants interacted with an ITS designed to teach a multi-step process for addressing constraint satisfaction problems to computer science undergraduates via an interactive simulation. This ITS provided hints to the student on how to improve their performance based on their behavior in the problem-solving task. When presented with a hint, participants were given the option to have the hint explained and were prompted to select the type of explanation wanted: "why" the system provided the hint (with an explanation of the user behavior triggering the hint) or "how" the system chose the hint to deliver (with an explanation of the algorithms reasoning for the hint). Fifty-four percent of all participants selected the option to receive an explanation along with the hint. Of these participants, close to 70% wanted to know why the system was providing the hint it chose. Participant opinions were also collected in a post-experiment survey, which Putnam and Conati (2019) report provide insight into how the users use this type of feedback to adjust their learning behavior. This is consistent with Johnson & Priest (2014)'s

Feedback Principle of Multimedia Learning, stating that novice participants learn better with explanatory feedback than corrective feedback alone.

These results of user opinions reported in Putnam and Conati (2019)'s pilot study satisfy Kirkpatrick and Kirkpatrick (2016) first level of training effectiveness, indicating participants found the training engaging and the transparency interventions to be favorable. This study sets the stage to explore the use of transparency interventions in the ITS domain on Kirkpatrick & Kirkpatrick (2016) Level 2 of training effectiveness which concerns the degree to which participants acquire the intended skills and knowledge of the training. Kirkpatrick's (2016) Four-Level Training Evaluation Model can be used to evaluate the effectiveness of training based on user Reaction, Learning, Behavior, Results. Levels 3 and 4 of Kirkpatrick's model concern the application of training to on-the-job tasks and outcomes. Although important, these levels do not apply to the current study, which aims to expand on the research of Putnam & Conati (2019) by examining the effects of transparency on the learning outcomes of an intelligent tutoring system. Putnam and Conati (2019) satisfies the first level of training effectiveness, Reaction, indicating users found it enjoyable and preferred receiving explanations. While it is important to gauge how users interact with a system and if they enjoy using it, we know from the training literature that preference is not always indicative of improved performance (Pashler et al, 2008; Yan et al, 2016; Rhodes, et al 2020) . The current study sets the stage to explore the use of transparency interventions in the ITS domain on Kirkpatrick and Kirkpatrick (2016) Level 2 of training effectiveness, Learning, which concerns the degree to which participants acquire the intended skills and knowledge of the training. Often, users are unaware of the components of a system that will have an effect on their task performance. This phenomenon is well observed in learning science literature, where people typically are poor at assessing their own level of knowledge and

determining not-yet-learned content (Pashler et al., 2008; Yan et al., 2016; Kirschner, 2017; Rhodes et al., 2020).

The Current Study

The first goal of current study is to examine whether adding transparency into an adaptive training system produces greater learning performance in an assembly task than when transparency is absent. This aim serves to fulfil Kirkpatrick and Kirkpatrick's (2016) second level of training effectiveness concerning the degree to which participants acquire skills and knowledge by providing transparency of the systems reasoning for the feedback it is giving. The second goal of this study is to assess how system transparency can affect user perceptions of their own knowledge and perceptions of the system itself. To examine both of these research goals empirically, adaptivity was implemented through assigning error-sensitive feedback and assigned restudy of missed concepts. Transparency was manipulated through this error-sensitive feedback, which was delivered as a text-based reasoning for assigning participants to restudy concepts they had missed. This form of transparency, as opposed to the other methods discussed above, was considered to fit the task as all three components of transparency in Chen et al's (2014) model in a way that was easily interpretable to the learner.

Specific hypotheses include:

1. Based on the evidence supporting the use of adaptive instruction over traditional instruction, (e.g., Metzler-Baddeley & Baddeley, 2009; Mettler et al., 2016) participants are predicted to show greater learning performance in the adaptive instruction conditions than in the static instruction conditions.

2. Learning performance is predicted to be higher in the transparent conditions than in the non-transparent conditions due to the performance benefit of automation transparency observed in other task domains (Bhaskara et al., 2021; Sargent et al., 2023)
3. As observed in Putnam and Conati (2019), transparency will produce more positive perceptions of training.

METHODS

Participants.

60 (32 Male, 27 Female, 1 Non-Binary) Undergraduate Psychology students from Colorado State University participated in this study online for partial, optional course credit. Of the sample, 97% were between ages 18 to 24, 2% between 25-34, and 2% were over age fifty-five. A priori power analysis conducted using G*Power (Faul et al., 2007) indicated that a sample size of 73 participants would be necessary to detect an effect size of 0.5 at alpha.05 with 80% power.

Task.

Participants were instructed on an assembling task for 8 total shapes. They studied the shapes passively by watching a 30 second assembly video, then were instructed on three learning objectives: the sequence of the bars to create each shape, the attachment point for each piece, and their placement. They received feedback based on the condition they were assigned and were assigned restudy based on condition. Adaptive conditions were assigned restudy based on their own performance, while static conditions were assigned restudy based on another, yoked participant's, study schedule. Table 1 illustrates the feedback given for each condition. Once the training phase was complete, participants were tested on all three learning objectives for each shape.

Materials.

Stimuli. A set of 8 shapes adapted from (Clegg et al., 2022) consisting of 4 to 9 components that attach as 3 to 8 possible positions to other parts. Figure 1 shows an example of one shape.

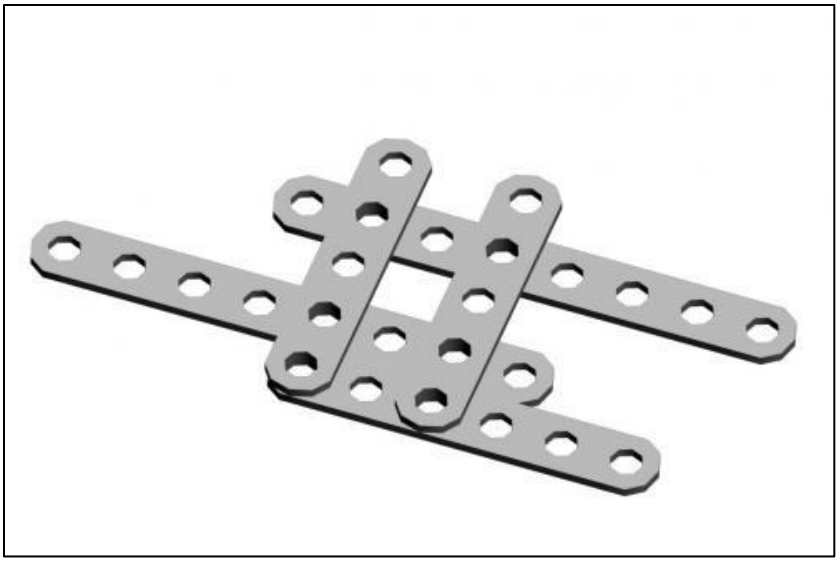

Figure 1. Example of a shape that participants will be instructed to build.

Participant perceptions questionnaire. Participants completed a short questionnaire related to their perceptions of the training and whether they noticed the system adapting to their performance. This also assessed participants ratings of their own learning and their likelihood to remember how to build each shape.

Procedure.

Participants accessed the experiment through a Qualtrics survey on a computer. Once they provided consent to participate, they were instructed that their task is to learn how to assemble 8 shapes and would be tested on how well they could remember the sequence order in which part is presented, which holes it attaches with, and where it goes in the shape. Instructional content was presented in blocks of a passive phase, active phase, and testing phase for 2 objects per block. Participants began each block in the passive learning phase, where they viewed each of the assembly videos in a random order. Next, participants entered the training phase with either static or adaptive instruction, determined the type of restudy intervention they receive. In the adaptive instruction conditions, participants were assigned restudy of the questions they answered incorrectly during training and in the static instruction conditions, they were assigned

restudy based on an adaptive participant's restudy schedule (see Table 3 for an explanation of the yoked design). Participants in transparent conditions were provided with feedback explaining the system's decision to assign restudy. while non-transparent conditions received generic feedback. As mentioned, transparency was implemented in the form of text explanations of the system's decision to assign restudy of concepts. Table 1 shows examples of feedback in each of the four conditions. Once the training phase was complete, they were tested on the three learning objectives (order of assembly materials, attachment points, and placement) without feedback. Figure 2 provides examples of each of these learning objectives.

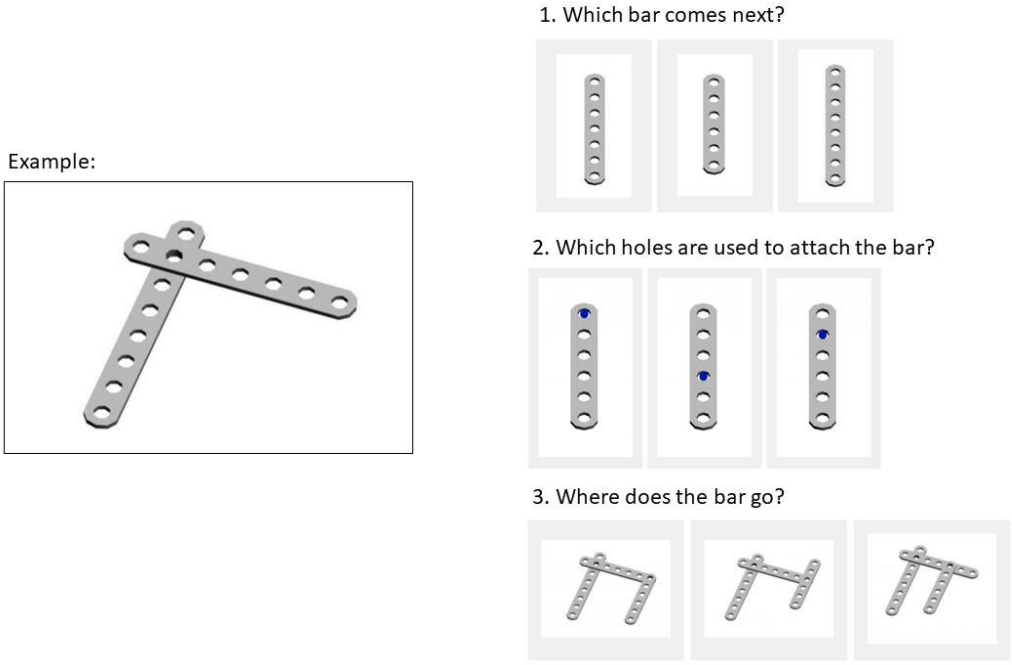

Figure 2. Example of an incomplete shape and the items for each learning objective that participants were shown. Question 1 captures the learning objective of order of assembly, question 2 captures attachment points, and question 3 captures placement.

Study Design.

The proposed study utilized a 2 (Instruction: adaptive or non-adaptive) X 2 (Transparency: present or absent) between-subjects yoked design to assess differences in learned content.

Participants were randomly assigned to either an adaptive or non-adaptive instruction instructional system with either transparent feedback related to the system's assessment of the user or standard performance feedback with no transparency. The order in which the shapes were presented was randomized for each participant.

Table 1. Between-subjects experimental design

		Instruction Type	
		Adaptive	**Static**
Transparency	**Transparency**	Adaptive Instruction w/ Transparency	Static Instruction w/ Transparency
	No-Transparency	Adaptive Instruction w/o Transparency	Static Instruction w/o Transparency

Table 2. Examples of feedback given to each condition.

Adaptive X Transparent Feedback *"You responded to 2 of 3 questions incorrectly. Your areas for improvement are Attachment Points and Order of Assembly of each shape."*	Static X Transparent Feedback *"The system has assigned restudy for generally difficult concepts to improve your performance on the test."*
Adaptive X Non-Transparent Feedback *"You will now restudy areas that need improvement."*	Static X Non-Transparent Feedback *"You will now restudy areas that are generally difficult."*

Yoking Conditions. To reduce the confound of restudy, participants in the static condition were assigned to the same number of presentations as another participant in the adaptive condition. The first 30 participants were run in the adaptive conditions, because their sequences provided the yoking input for the remaining 30 participants in the static conditions. Table 4 describes how each participant was assigned the study schedule.

Table 3. Example of the experimental schedule for yoking conditions

Data Collection Round 1		Data Collection Round 2	
Adaptive Condition	Assigned Restudy	Static Condition	Assigned Restudy
P1	Performance based	P5	Same as P1
P2	Performance based	P6	Same as P2
P3	Performance based	P7	Same as P3
P4	Performance based	P8	Same as P4

RESULTS

Restudying.

Participants in the adaptive conditions were assigned an average of 12.60 (*SE* = 1.23) items to restudy out of 48 possible items. In examining whether transparent feedback could lead to fewer cases of restudy, an Independent Samples t-test revealed no significant difference in the amount of items studies between the adaptive conditions ($t(29) = -1.20$, $p = 0.24$, $d = -0.43$). Participants in the Adaptive Transparency group restudied an average of 14 (*SE* = 1.62))items, while those in the Adaptive Non-Transparency group restudied an average of 11 (*SE* = 1.62) items. Due to the yoked design, participants assigned to static instruction did not differ from the adaptive conditions in the number of items restudied.

Performance in Early versus Late Shape Construction.

As mentioned above, participants were instructed on 3 Early-stage and 3 Late-stage steps for assembling each shape. For all 8 shapes, this amounts to 24 items in each category. Overall, participants correctly responded to an average of 20.60 (*SE* = 0.40) Early-stage items and 21.3 (*SE* = 0.41) Late-stage items.

Early-stage item performance. A 2(Instruction Type) by 2(Transparency Condition) conducted on Early-stage item test scores revealed no main effect of instruction type ($F(1,56) < 1$) nor transparency ($F(1,56) < 1$). There was also no significant interaction between both variables. Group means and standard errors are reported in Table 4.

Table 4.
Means and standard deviations for Ealy-Stage Item Score as a function of a 2(Adaptive Condition) X 2(Transparent Condition) design. M and *SE* represent mean and standard error, respectively.

	Transparency Condition			
	Non-transparent		Transparent	
Adaptive Condition	*M*	*SE*	*M*	*SE*
Adaptive	21.00	0.63	20.27	1.16
Static	20.73	0.69	20.20	0.70

Late-stage item performance. Participants in the Adaptive Non-transparency condition and static A 2 (Instruction Type) by 2 (Transparency Condition) conducted on revealed no main effect of instruction type ($F(1,56) < 1$) nor transparency ($F(1,56) < 1$). There was also no significant interaction between both variables. Group means and standard errors are reported in Table 5.

Table 5.
Means and standard deviations for Late-Stage Item Score as a function of a 2 (Adaptive Condition) X 2 (Transparent Condition) design. M and *SE* represent mean and standard error, respectively.

	Transparency Conditions			
	Non-transparent		Transparent	
Adaptive Condition	*M*	*SE*	*M*	*SE*
Adaptive	21.53	0.66	20.60	1.24
Static	21.47	0.53	21.40	0.72

Overall Learning Performance.

Number of shapes correctly built. Participants were scored on the number of shapes they were able to correctly build on the test. The test consisted of 6 items examining the steps in the build process they had received training on. Performance was assessed by whether participants responded to all 6 items for each shape correctly or had one or more errors. The average amounts of shapes correctly built at test for each condition are reported in Table 6. No main effect of

adaptative condition ($F(1,56) < 1$), nor transparent condition ($F(1,56) < 1$) was observed. There was also no significant interaction between adaptive condition and transparency on the number of shapes correctly built ($F(1,56) < 1$). Planned contrasts revealed a non-significant difference between instruction types in non-transparent conditions ($t(28) = 0.66$, $p = 0.52$, $d = 0.21$) as well as in the transparent conditions ($t(28) = 0.33$, $p = .75$, $d = 0.14$), indicating that regardless of whether transparency was present or absent, instruction type did not significantly affect the number of shapes correctly built. These results indicate that H1, predicting higher learning performance in transparent conditions was not supported.

Table 6.
Means and standard errors for Number of Shapes Correctly Built as a function of a 2(Adaptive Condition) X 2(Transparent Condition) design. M and *SE* represent mean and standard error, respectively.

	Transparent Condition			
	Non-transparent		Transparent	
Adaptive Condition	*M*	*SE*	*M*	*SE*
Adaptive	5.27	0.48	4.93	0.60
Static	4.80	0.52	4.67	0.55

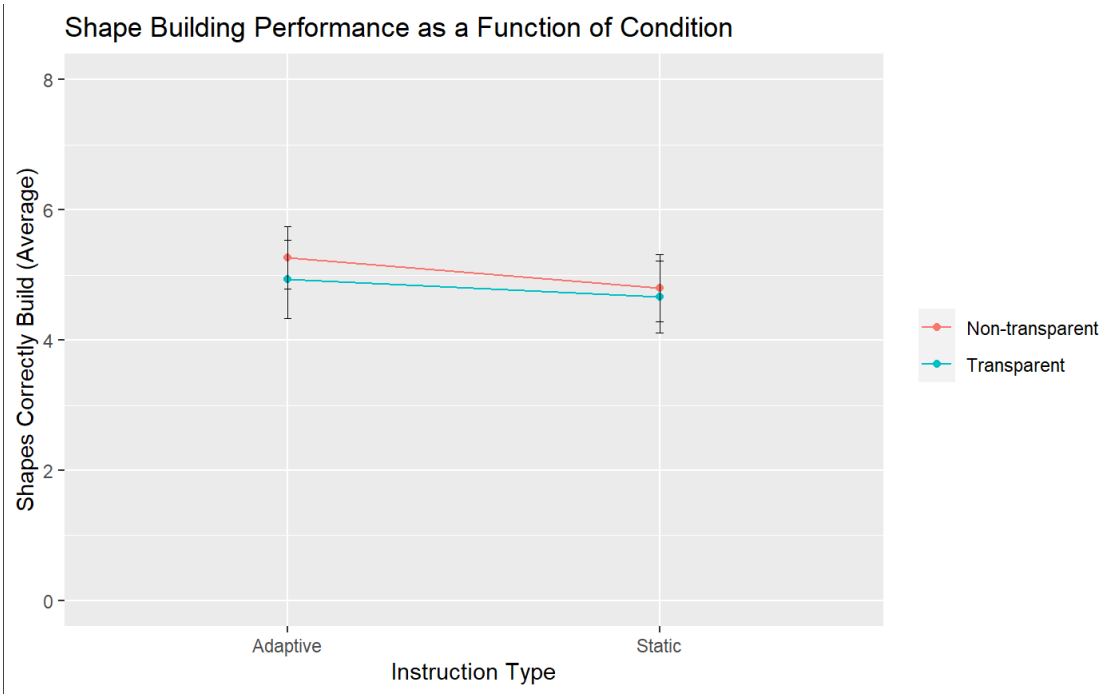

Figure 3. Shape Building Performance as a function of condition.

Improvement from pre-test to post-test. To analyze the change in performance from initial training performance to test performance, participant gain scores were calculated using the following formula: ([Post-Test Score – Pre-Test Score]/[Total Score – Pre-Test Score]). Analyzing gain scores normalizes learning performance to account for the participant's scores at pre-test and room for improvement at test. Average gain scores between conditions are reported in Table 7. in A 2 (Adaptive Condition) by 2 (Transparency condition) ANOVA was used to determine if there were any significant differences in learning gains. There was no main effect of adaptive condition ($F(1,56) < 1$), nor was there a main effect of transparent condition ($F(1,56) < 1$). There was also no significant interaction between adaptive and transparent interventions ($F(1,56) = 2.62$, $p = 0.11$, $\eta 2 = 0.04$). These results indicate that H2, predicting higher learning performance in transparent conditions was not supported.

Table 7.

Means and standard deviations for Gain Score as a function of a 2(Adaptive Condition) X 2(Transparent Condition) design. M and *SE* represent mean and standard error, respectively.

	Transparency Condition			
	Non-transparent		Transparent	
Adaptive Condition	*M*	*SE*	*M*	*SE*
Adaptive	36.96	9.91	56.60	8.91
Static	43.33	7.61	33.63	9.65

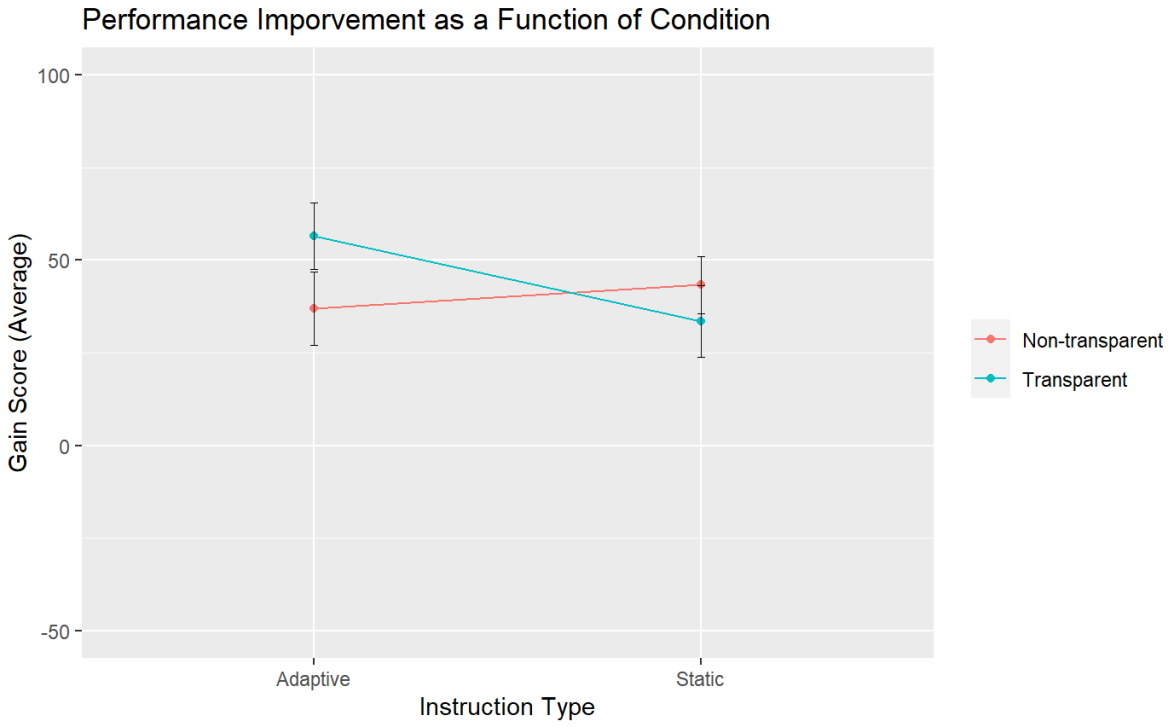

Figure 4. Performance Improvement as a function of Condition

Interestingly, when removing the participants assigned to conditions that received non-transparent feedback, a one-tailed independent samples t-test revealed that the impact of instruction type on learning gains was significant $t(28) = -1.75$, $p = .05$, $d = -0.64$) . This trend was not observed in the non-transparent conditions $t(28 = 0.51$, $p = .61$, $d = 0.19)$, this shows a

positive influence of adaptivity on learning gains, only when participants are given some indication that the system is in fact adapting to their performance.

Performance in Learning Objectives.

Order. Isolating test performance by the 16 test items related to the learning objective of Order (selecting the bar that came next in the sequence), participants in the Adaptive Transparency condition correctly responded to an average of (*SE* = 1.02) items, Adaptive Non-transparency averaged 13.7 (*SE* =0.70), Static Transparency averaged 13.5 (*SE* = 0.77), and Static Non-transparency averaged 13.70 (*SE* = 0.49). A between-subjects ANOVA showed no significant main effects of transparency ($F(1,56) < 1$), nor instruction type ($F(1,56) < 1$) and no significant interaction ($F(1,56) < 1$).

Attachment Point. Isolating test performance by the 16 test items related to the learning objective of Attachment Point (selecting the correct hole the bar uses to attach), participants in the Adaptive Transparency condition correctly responded to an average of 13.9 (*SE* = 0.73) items, Adaptive Non-transparency averaged 13.9 (*SE* =0.61), Static Transparency averaged 14.2 (*SE* = 0.47), and Static Non-transparency 14.30 (*SE* = 0.37). A between-subjects ANOVA showed no significant main effects of transparency ($F(1,56) < 1$), nor instruction type ($F(1,56) < 1$) and no significant interaction ($F(1,56) < 1$).

Location. Isolating test performance by the 16 test items related to the learning objective of Location (where the bar is placed on the shape), participants in the Adaptive Transparency condition correctly responded to an average of 14 (*SE* = 0.78) items, Adaptive Non-transparency averaged 15.10 (*SE* =0.41), Static Transparency averaged 14.3 (*SE* = 0.55), and Static Non-transparency 14.50 (*SE* = 0.41). A between-subjects ANOVA showed no significant main effects

of transparency ($F(1,56) < 1$), nor instruction type ($F(1,56) < 1$) and no significant interaction ($F(1,56) < 1$).

Metacognitive Predictions.

A 2(adaptive condition) by 2(transparency condition) ANOVA was used to determine if there were significant difference in participants' rated confidence (on a scale of 1 to 100) that they would remember the steps for assembly on the test. No main effect of either adaptive condition ($F(1,56) < 1$) nor transparency condition ($F(1,56) = 1.99$, $p = .16$, $\eta 2 = 0.0$). There was also no significant interaction ($F(1, 56) < 1$). Group means are reported in Table 8. Participants' predictions of their ability to remember were slightly correlated with their learning gains ($r = .26$, $p < .05$).

Table 8.

Means and standard deviations for Participant Confidence Ratings as a function of a 2(Adaptive Condition) X 2(Transparent Condition) design. M and *SD* represent mean and standard deviation, respectively.

	Transparent Condition			
	Non-transparent		Transparent	
Adaptive Condition	*M*	*SD*	*M*	*SD*
Adaptive	49.75	22.33	50.13	20.66
Static	52.18	20.30	52.63	25.92

Perceptions of Training.

Enjoyment. Participants were asked to rate their enjoyment of the training on a 7-point Likert scale, with 1 being "Strongly Disagree" and 7 being "Strongly Agree". Group means are reported in Table 9. There was no main effect of Adaptive condition ($F(1,56) < 1$) nor Transparency condition ($F(1,56) < 1$). There was also no significant interaction ($F(1,56) < 1$). This indicates

no meaningful difference in participants enjoyment of the training, regardless of whether it was tailored to their performance or if they were given transparent feedback.

Table 9.

Means and standard deviations for Participant Rated Enjoyment as a function of a 2(Adaptive Condition) X 2(Transparent Condition) design. M and *SD* represent mean and standard deviation, respectively.

	Transparent Condition			
	Non-transparent		Transparent	
Adaptive Condition	*M*	*SD*	*M*	*SD*
Adaptive	2.40	0.83	2.80	1.15
Static	2.33	0.72	2.60	0.91

Perceived difficulty of training. Participants were to indicate asked how strongly they agreed with the statement "The difficulty of the task was appropriate for my skill level" on a 7-point Likert Scale, with 1 being "Strongly Disagree" and 7 being "Strongly Agree". Group means are reported in Table 10. There was no main effect of Adaptive condition ($F(1,56) < 1$) nor Transparency condition ($F(1,56) < 1$). There was also no significant interaction ($F(1,56) < 1$).

Table 10

Means and standard deviations for difficulty as a function of a 2(Adaptive Condition) X 2(Transparent Condition) design. M and *SD* represent Mean and Standard Deviation, respectively

	Transparent Condition			
	Non-transparent		Transparent	
Adaptive Condition	*M*	*SD*	*M*	*SD*

Adaptive	2.27	0.46	2.73	0.96
Static	2.13	0.74	2.33	0.82

Noticing adaptation. To assess whether participants were able to detect when the system was assigning restudy based on their performance, they were asked to indicate how strongly they agreed with the statement "The system customized my training based on my answers" on a 7-point Likert Scale, with 1 being "Strongly Disagree" and 7 being "Strongly Agree". Group means are reported in Table 11. There was no main effect of Adaptive condition ($F(1,56) < 1$) nor Transparency condition ($F(1,56) < 1$). There was also no significant interaction ($F(1,56) < 1$).

Table 11

Means and standard deviations for noticing the customized aspect of the training as a function of a 2 (Adaptive Condition) X 2 (Transparent Condition) design. M and *SD* represent mean and standard deviation, respectively.

	Transparency Condition			
	Non-transparent		Transparent	
Adaptive Condition	*M*	*SD*	*M*	*SD*
Adaptive	2.67	1.05	2.20	0.94
Static	2.73	0.80	2.53	0.64

Perceived system accuracy. Participants also rated how accurately the system was able to assess their performance. Group means are reported in Table 12. There was a significant effect of both adaptive condition ($F(1,56) = 7.22$, $p < .01$, $\eta 2 = 0.11$) and transparency condition ($F(1.56) = 3.89$, $p < .05$, $\eta 2 = 0.11$). No significant interaction was found ($F(1,56) < 1$). The system was actually rated as the least accurate in the adaptive transparency condition, perhaps indicating that

increased transparency in adaptive instruction could cause an unwanted decrement to perceptions of accuracy.

Table 12.
Means and standard deviations for Participants' Perceived System Accuracy as a function of a 2 (Adaptive Condition) X 2 (Transparent Condition) design. Note. M and *SD* represent mean and standard deviation, respectively

	Transparent Condition			
	Non-transparent		Transparent	
Adaptive Condition	*M*	*SD*	*M*	*SD*
Adaptive	2.40	0.51	1.80	0.68
Static	2.67	0.98	2.53	0.64

.

DISCUSSION

The findings of the current study must be interpreted with a lack of statistical power in mind. Overall, there was not sufficient evidence to support the hypothesis that participants would show greater learning performance in the adaptive instruction conditions than in the static instruction conditions. This was indicated by the weak effects found between adaptive and static instruction in the number of shapes that participants were able to correctly build at test, and further supported by minimal differences in learning gains from pre-test to post-test. However, it is important to note that the effects of instruction type on participant learning gains was found to be significant in the transparent conditions. This could perhaps indicate that automated systems benefit from the inclusion of transparency, but the effects do not hold in traditional instructional systems. Indeed, if this effect of adaptivity is only present in transparent systems, it may indicate that adaptive instruction is most beneficial to learning when learners have an understanding of how the system is responding to their performance.

Although adaptive training has shown to be beneficial for skill acquisition, reviews of its impact on performance recommend careful consideration when determining the adaptive interventions implemented and their appropriateness for the task domain (Durlach & Lesgold, 2012). It is possible that assigning a single restudy attempt was not a sufficient adaptation for training in the shape building task. Moreover, the error-sensitive feedback provided to participants included a score and their areas of improvement. Perhaps participants would have benefitted from more specific feedback that provides hints or cues to help encoding during restudy. A similar explanation can be given to the weak effects of transparency. As shown in Bhaskara et al., (2021) there are multiple ways to increase system transparency, and it may be

that simply including a text explanation of the system's reasoning for assigning restudy was not helpful in directing participants' attention to the areas of improvement.

Participants' metacognitive predictions in their ability to remember each shape at test also remained unaffected by the type of instruction or amount of transparency they received in their feedback. However, there was a slight but significant correlation ($r = .26$, $p < .05$) between their predictions and their actual learning gains, indicating that they were somewhat able to predict their own performance at test based on their perceptions of learning from training.

When assessing participants' perceptions of training, there was no difference in the reported enjoyment of the training between conditions. However, some evidence was shown to support that adaptive instruction and transparency influence perceptions of how accurate the system is at understanding performance. Surprisingly, the adaptive conditions were rated as significantly less accurate in assessing participants' learning performance than the static condition. When transparency was added to these conditions, perception of accuracy also lowered by a marginally non-significant difference.

Taken together, the trends of these findings suggest that using transparency to explain the reasoning of adaptive interventions may actually be detrimental to a learner's perceptions of the system. This notion is in contrast to the findings reported in Putnam and Conati (2019), however, it is important to note that their pilot study did not utilize a yoked design to compare the perceptions of training between transparent and non-transparent conditions. In the case of intelligent tutoring systems, the yoked design allows us to draw conclusions about whether tailoring transparency and adaptive manipulation to the individual are the reasons for change in performance, rather than resulting from the greater amount of practice that can occur. Perhaps the Putnam and Conati (2019) study fulfilled Kirkpatrick's first level of training effectiveness

because all participants were given training interventions that were tailored to their own skill and not another participant's, but we cannot attribute the enjoyment of the training to being tailored without a yoked group to serve as a control. While the current study did not find sufficient evidence to demonstrate participant learning performance and, in turn, satisfy Kirkpatrick's guidelines on training effectiveness, more evidence is still needed to confirm whether the use of automation transparency increases training outcomes.

Limitations.

It must be noted that this study had multiple limitations, the first being the scope of adaptive instructional interventions as well as methods of implementing transparency. As previously discussed, there are many ways to adapt instruction to a learner's skill level, and this experiment only examine assigning restudy. While participants in the Adaptive Transparency condition did receive scores in their feedback, perhaps performance in this task domain would increase over static instruction with the inclusion of more specific error-sensitive feedback. Moreover, the algorithm for determining feedback and restudy was relatively simplistic, and the benefits of easy-to-understand automation transparency may be more salient complex systems. In addition to these limitations, due to inconsistent enrollment for the study, the sample was insufficiently powered. Future studies should examine multiple types of adaptive interventions and the effects of different categories of automation transparency.

Much of the automation literature examines the differences in transparency effects when automation fails or succeeds in its assessment of condition and intervention. In a learning context, automation failure could be represented as an incorrect assessment of the learner's training performance. For example, if an ITS provides a student with a score and explanation of the need for restudy that does not match the students' actual performance, will the student

notice? Sargent et al. (2023) find in their meta-analysis that transparency's performance benefits are most pronounced in failure performance, or when automation fails and the user is able to notice and correct course. Real-world automated systems, including ITS, cannot be perfectly reliable. As these systems are prone to errors, such as, for example, making an incorrect inference regarding user learning level, it will be critical to understand what this means for students using these tools.

CONCLUSION

An important takeaway from this study is that adaptive training and automation transparency describe broad areas of research. Careful consideration must be taken to ensure the interventions used are the most appropriate for the task domain. While adding transparency to a training system may not always affect learning performance, it may have the potential to create negative perceptions of the intelligent tutoring system. Although reviews of empirical studies have shown a performance benefit from transparent systems, more research is needed on parameters that may determine whether transparency hurts or helps task performance. Furthermore, implementation of the correct methodology when studying adaptive instruction and automation transparency is vital to answer the questions still present in these domains.

REFERENCES

Astwood, R. S., Van Buskirk, W. L., Cornejo, J. M., & Dalton, J. (2008, September). The impact of different feedback types on decision-making in simulation-based training environments. *Proceedings of the human factors and ergonomics society annual meeting*, 52(26), 2062-2066

Bhaskara, A., Duong, L., Brooks, J., Li, R., McInerney, R., Skinner, M., ... & Loft, S. (2021). Effect of automation transparency in the management of multiple unmanned vehicles. *Applied Ergonomics, 90*, 103-243.

Bhaskara, A., Skinner, M., & Loft, S. (2020). Agent transparency: A review of current theory and evidence. *IEEE Transactions on Human-Machine Systems, 50*(3), 215-224.

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). Situation awareness-based agent transparency. Army research lab aberdeen proving ground md human research and engineering directorate.

Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling.

Durlach, P. J. (2019). Fundamentals, flavors, and foibles of adaptive instructional systems. *Adaptive Instructional Systems: First International Conference Proceedings 2*, 76-95.

Durlach, P. J., & Lesgold, A. M. (Eds.). (2012). Adaptive technologies for training and education. Cambridge University Press.

Durlach, P. J., & Ray, J. M. (2011). Designing adaptive instructional environments: Insights from empirical evidence. US Army Research Institute for the Behavioral and Social Sciences.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human factors, 37(*1), 65-84.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175-191.

Johnson, C. I., Bailey, S. K., & Van Buskirk, W. L. (2017). Designing effective feedback messages in serious games and simulations: A research review. *Instructional techniques to facilitate learning and motivation of serious games*, 119-140.

Johnson, C. I., Priest, H. A., Glerum, D. R., & Serge, S. R. (2013). Timing of feedback delivery in game-based training. *Proceedings of the Interservice/Industry Training, Simulation & Education Conference, Orlando, FL, 2013*. Arlington, VA: National Training Systems Association.

Johnson, C. I., & Priest, H. A. (2014). 19 The Feedback Principle in Multimedia Learning. *The Cambridge handbook of multimedia learning, 449.*

Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). Kirkpatrick's four levels of training evaluation. Association for Talent Development.

Kirschner, P. A. (2017). Stop propagating the learning styles myth. *Computers & Education, 106*, 166-171.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, *119*(2), 254.

Landsberg, C. R., Astwood Jr, R. S., Van Buskirk, W. L., Townsend, L. N., Steinhauser, N. B., & Mercado, A. D. (2012). Review of adaptive training system techniques. *Military Psychology, 24*(2), 96-113.

Mayer, R. E. (2014). Multimedia instruction. *Handbook of research on educational communications and technology*, 385-399.

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology: General*, *145*(7), 897.

Mettler, E., Burke, T., Massey, C. M., & Kellman, P. J. (2020). Comparing adaptive and random spacing schedules during learning to mastery criteria. In *CogSci... Annual Conference of the Cognitive Science Society. Cognitive Science Society (US). Conference*, 2020, 773

Metzler-Baddeley, C., & Baddeley, R. J. (2009). Does adaptive training work?. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 23*(2), 254-266.

Nwana, H. S. (1990). Intelligent tutoring systems: an overview. *Artificial Intelligence Review, 4*(4), 251-277.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*(3), 286-297.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological science in the public interest, 9*(3), 105-119.

Pon-Barry, H., Schultz, K., Bratt, E. O., Clark, B., & Peters, S. (2006). Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education, 16*, 171 – 194.

Putnam, V., & Conati, C. (2019, March). Exploring the Need for Explainable Artificial Intelligence (XAI) in Intelligent Tutoring Systems (ITS). *IUI Workshops,19*(2019).

Rhodes, M. G., Cleary, A. M., & DeLosh, E. L. (2020). A guide to effective studying and learning: Practical strategies from the science of learning. Oxford University Press.

Sargent, R., Walters, B. & Wickens, C. (2023) Meta-Analysis Qualifying and Quantifying the Benefits of Automation Transparency to Enhance Models of Human Performance. *Proceedings HCI-International.* Copenhagen Denmark.

Smith, S. G., & Sherwood, B. A. (1976). Educational uses of the PLATO computer system. *Science, 192*, 344–352.

Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction, 8*(4), 351-362.

Schwonke, R., Hauser, S., Nückles, M., & Renkl, A. (2006). Enhancing computer-supported writing of learning protocols by adaptive prompts. *Computers in Human Behavior, 22*, 77-92.

Tseng, J. C. R., Chu, H., Hwang, G., & Tsai, C. (2008). Development of an adaptive learning system with two sources of personalization information. *Computers & Education, 51*, 776-786.

U.S. Department of Education, National Center for Education Statistics (2022). *Impact of the Coronavirus (COVID-19) Pandemic on Public and Private Elementary and Secondary Education in the United States (Preliminary Data): Results from the 2020-21 National Teacher and Principal Survey (NTPS).*

Wagman, M. (1980). PLATO DCS: An interactive computer system for personal counseling. *Journal of Counseling Psychology, 27*(1), 16–30. https://doi.org/10.1037/0022-0167.27.1.16

Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2022). *Engineering psychology and human performance.* Routledge.

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General, 145*(7), 918–933. https://doi.org/10.1037/xge0000177

Yazdani, M. (1986). Intelligent tutoring systems: an overview. *Expert systems, 3*(3), 154-163.

APPENDICIES

Appendix A: Post-survey training perceptions questionnaire

Please respond to the following items with your level of agreement for each statement, with 1 being "Strongly Disagree" and 7 being "Strongly Agree."

1. Overall, I liked the content in this training.
2. I believe that the feedback I received focused my attention on learning strategies to perform this task better.
3. The difficulty of questions was appropriate for my skill level.
4. The system customized my training based on my answers.
5. The system was able to accurately assess my performance.
6. If given the opportunity, I would restudy more items before the test.

Appendix B: Demographics Questionnaire

1. What is your major? _____
2. Which best describes your academic performance?
    a. Freshman
    b. Sophomore
    c. Junior
    d. Senior
    e. Graduate Student
3. Which age range do you fall into?
    a. 18 to 24
    b. 25 to 34
    c. 35 to 44
    d. 45 to 54
    e. 55+
4. How would you describe your gender?
    a. Man (including cisgender and transgender men)
    b. Woman (including cisgender and transgender women)
    c. Non-binary
    d. Other (please specify)
    e. Prefer not to say.