A Systematic Review of Literature on the Effectiveness of Intelligent Tutoring Systems in STEM

Shi Feng
Computer and Information Technology
Purdue University
West Lafayette, U.S.A.
https://orcid.org/0000-0003-0675-6450

Alejandra J. Magana
Computer and Information Technology
Purdue University
West Lafayette, U.S.A.
https://orcid.org/0000-0001-6117-7502

Dominic Kao
Computer and Information Technology
Purdue University
West Lafayette, U.S.A.
https://orcid.org/0000-0002-7732-6258

***Abstract*—Intelligent tutoring systems (ITS) have shown to be useful learning aids for helping students learn STEM subjects. Previous studies on ITS tend to focus on developmental aspects of the system, such as system design, programming architecture, and dialogue moves. In this systemic literature review, we focus on pedagogical aspects of ITS within STEM domains. Specifically, we identified the implemented scaffolding approach and the grounding on learning theories of ITS implementations. Specific research questions were: (1) what types of knowledge (i.e., conceptual learning, problem-solving, and model building) are delivered via an ITS within STEM domains? (2) what pedagogies or scaffolding methods are used to guide the ITS learning experiences? (3) what are the characteristics of the research designs and specific learning outcomes when learning with the ITS? The steps followed for performing this systematic literature review were: (1) identifying the scope and research questions, (2) defining the inclusion and exclusion search criteria of literature, and (3) classifying and cataloging the literature sources that use ITS for STEM in classroom research. The final data set is comprised of a total of 22 papers that meet our criteria. We found a lack of fine-grained research on the effectiveness of using ITS to improve the three major learning modes: conceptual learning, problem-solving, and model building, particularly in STEM domains. In addition, we recommend that research conducted on ITS and other learning technology aids should emphasize the utilization of well-established learning theories and pedagogical scaffolding methods so that ITS will be more accessible to STEM educators for introducing ITS to their students to better learn STEM subjects.**

Keywords—intelligent tutoring systems, STEM, systematic literature review

I. INTRODUCTION

Currently, according to the U.S. Education Department's National Center for Education Statistics [1], 50% of STEM undergraduate students are at risk of leaving the field before completing a college degree. Therefore, there is an urgency for technology-based learning assistants tailored to the particular needs of individual students in STEM majors. Computer tutoring systems that are deemed as "intelligent" enough to meet the standards of expert human tutors have become a reality[2]. The current study draws from the previous meta-analysis literature to conduct a systematic review of the learning theories, scaffolding methods, and effectiveness researched on the available ITS within the past two decades. Specifically, we are interested in STEM domains with a focus on conceptual learning, problem-solving, and model building using an intelligent tutoring system.

II. BACKGROUND

An intelligent tutoring system (ITS) is generally defined as a computer learning environment that helps students master knowledge by implementing artificial intelligent algorithms. Such algorithms are usually adapted to the individual student's mastery level and unique learning types in presenting content, asking questions, and giving feedback [3]. More recently, ITS can now engage in conversational dialogues using natural language between an animated talking head agent and the human learner [4]. This advancement would allow for the next generation of ITS to mimic human speech, facial expressions, and even gestures.

The main goal of building an ITS is to simulate a human tutoring environment [5]. The basic architecture of these systems follows the principles from classical learning theories such as social and cognitive constructivism, where the students learn by discovering and assimilating new information into their existing knowledge structure and interacting with their peers [6]. In order to provide meaningful instruction and scaffolding to the students, ITS need to adapt the appropriate content and presentation with respect to the students' zone of proximal development (ZPD) [7]. ITS can do this by actively sequencing the students' mastery levels and monitoring the students' affective states to ensure engagement so they will not become bored and therefore "zoning out" [8].

ITS incorporates modern learning theories such as the cognitive-affective theory of multimedia learning. Students are being put into a virtual reality environment with dynamic multimodal instructional materials that aim to motivate the student to use different sensory channels to process new information [9]. In addition, many of the features of modern ITS attempt to promote cognitive disequilibrium, where the system deliberately presents students with contradicting or false information, so the students acquire deep knowledge by resolving their state of confusion in complex learning [10].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

A. Types of Intelligent Tutoring Systems and Scaffolding Strategies

A core ambition of modern ITS is to engage students with complex learning, which would utilize several scaffolding strategies by combining the learning theories and the prescribed pedagogical methods. The currently most widely used type, the expectation and misconception tailoring type of ITS uses tutoring dialogues in natural language to tackle students' topic knowledge and identify their misconception that the intelligent tutor can then address and correct. AutoTutor is one such ITS that uses natural language dialogues to teach students STEM subjects such as physics [3]. A specific scaffolding method for this type of ITS using expectation and misconception has been developed both from the explanation-based constructivist theories of learning [11, 12], which previous research has suggested to be prevalent amongst human tutoring sessions [13]. AutoTutor, for example, uses a specific set of dialogue moves that tackle knowledge from general and shallow to deep and specific. It does so by using the main question (a general open-ended conceptual question), hints (giving a clue that guides that student in the right direction of the expected answer), and prompts (a final clue that will guide the student to find the specific concept missing from the previously made answers) [14]. These dialogues will often compel the students to reiterate their answers in many different ways until the system is confident that the student has identified all core parts of the knowledge to be learned.

Another type of ITS uses the constraint-based modeling (CBM) scaffolding approach, based on Ohlsson's model of learning from performance errors [15]. These ITS have a predefined set of problems and ideal solutions written by a human expert who is then represented in the system as a series of logical constraints. These constraints are then related to information that will give rise to a solution. The system identifies any error from the students from any violation of these logical constraints by not meeting the satisfactory answer. When the students make an error, a feedback message will be given to the students. Thus, the students will acquire the necessary knowledge by learning from errors. At the same time, the ITS can represent the students' knowledge as a set of violated and satisfactory constraints. The advantage with CBM is tutoring more interactive activity-based topics, such as database design in the case of the ITS KERMIT [16].

The last two major types of ITS deal directly with student modeling as its core scaffolding strategy: model tracing and Bayesian network modeling—the latter builds upon the former. Model-tracing involves writing a set of production rules that will map and model students' problem-solving processes. One of the most well-known cognitive models is the ACT-R model, which is the theoretical basis for ITS, such as the cognitive tutor [17]. ART-R represents declarative and procedural knowledge during learning via chunks (i.e., knowledge needed for solving a problem) and production rules (i.e., specified conditions and actions involved in solving a problem). The Cognitive Tutor uses this framework by creating different modules which track students' knowledge states. If an error is detected, immediate feedback is then given [17]. More recently, the theoretical frameworks of Bayesian network modeling are applied to student modeling that introduces probabilities of knowledge representation. Based on students' performance on solving a problem, the ITS will use the Bayesian network to predict students' further performances, whereby giving students hints and feedback to correct their misconceptions and potential future misconceptions.

B. Evidence of Learning with Intelligent Tutoring Systems

An early review of modern ITS systems showed an effect size of d= 1.0 when compared to the no tutoring condition [18]. Modern expectation misconception tailoring step-based tutoring systems using dialogue for scaffolding perform just as well as human tutors on STEM topics. An interesting finding is that the decreasing of granularity plateaus to a certain point relative to an increase in effect size in learning. ITS that is sub-step based rather than step based did not perform as well step based ITS, suggesting that there is an optimal granularity for maximizing learning using a step-based tutoring system that needs to be considered when designing ITS. However, most of the ITS performances were not up to par with expert human tutors, which resulted in a learning gain of d=2.0.

Ma, Adesope, Nesbit, & Liu [19] conducted a much more comprehensive meta-analysis of ITS of all knowledge domains and levels of education. The meta-analysis replicated the previous finding that ITS significantly improved learning relative to no tutoring but also other forms of instruction such as classroom learning. However, ITS did not help to learn significantly better than small group learning with a human tutor or one-on-one human tutoring. The use of ITS produced significant effect sizes in learning for all grade levels, with middle school and college students having larger effect sizes and elementary school or high school students. ITS produced moderate (g = .35 to .59) effect sizes for most STEM subjects. In addition, ITS helped students with low and medium prior knowledge but not high prior knowledge. This may be due to low samples in the high prior knowledge studies. Finally, it was found that ITS produced significant learning outcomes in both classroom and laboratory studies, with classroom learning having a marginally significant effect than laboratory learning.

Kulik and Fletcher [20] conducted a more rigorous meta-analysis on ITS that delved deeper into types of experiments conducted, particularly the appropriate usage of control groups and assessment instruments used for evaluating ITS effectiveness as a learning aid. They found that overall, ITS had moderate to large effects on improving learning. However, the effectiveness varies in accordance to the following: (1) when the test instruments were locally constructed rather than using standardized test items, ITS had a substantially larger effect size in learning, (2) when a large sample size was used in the evaluations, the effect size shrinks, (3) when participants were either of lower grade levels or were unfamiliar with using the system, the effect sizes were smaller, but can improve substantially over a long period of usage, and (4) ITS had significantly less effect on learning outcomes when multiple-choice tests were used to measure outcomes. In addition, Kulik and Fletcher [20] found that when the ITS Cognitive Tutor was used in the meta-analysis, the effect size decreased. This is likely due to the large-scale nature of the evaluation used in Cognitive Tutor, and standardized test items used almost exclusively for the system.

Based on the previous research, it is reasonable to assume that using ITS does contribute to better learning outcomes. However, what is missing in the current evaluation of ITS, and learning technologies in general, is the role of learning theories and specific scaffolding methods or pedagogy used for improving learning. The previous meta-analyses did not investigate further the specific production rules or scaffolding paths used in each of the systems. Similarly, while previous research identified ITS that used different implementation of knowledge representation and decision making to tailor scaffolding to individuals [19], they failed to identify the learning theories used to design the system in the first place. The current systemic review aims to address these questions in order to fill the gap of knowledge for designing and implementing tutoring systems for further students. We concentrate on STEM learning students using ITS teaching a STEM subject. We restrict our analysis on intelligent tutoring systems within the past two decades to give a more recent, up-to-date finding on the current development of learning technologies.

III. Methods

The steps used to obtain the literature for the current review were: (1) identifying the scope and research questions, (2) defining the inclusion and exclusion search criteria of literature, and (3) classifying and cataloging the literature sources that use ITS for STEM in classroom research. Since our primary research question is to investigate whether different types of pedagogical approaches and theories of learning will have an impact on conceptual learning, problem-solving, and model building via ITS, we primarily will restrict the literature search that includes three core elements in the papers—description of pedagogy used, theories of learning from which the pedagogy was derived, and experimental design that had directly investigated learning outcomes for the target population.

A. Identifying Inclusion/Exclusion Criteria

The following search criteria were used in the current study to identify literature that are pertinent to the research questions.

- The study has been published in an English peer-reviewed journal in the past 20 years.
- The intelligent tutoring system in question must be one built for a STEM subject for students learning the fundamental basics of the topic in STEM.
- The study must have utilized an empirical, experimental design to assess learning outcomes with at least one experimental treatment group and one control group for comparison.
- The study must have specified the scaffolding strategy frameworks and learning theory used for the system's architecture.
- The ITS must have built-in capabilities designed to improve conceptual learning, problem-solving, or model building, or a combination of the three learning types.

B. Finding and Cataloging Sources

The keywords used in the search criteria were "intelligent tutoring systems" and "performance" or "improvement" and excluding keywords such as "emotion" or "regulation," as the current research questions are not interested in the affective aspect of the tutoring systems. The publication dates were set to the years 2000 and onwards. From there, papers were selected based on the title and abstracts for the search criterion that the system must be built for a STEM subject. The primary database used was PsycInfo which consists of studies in the psychological and learning sciences. The second database used was the Education Resources Information Center (ERIC) which consisted of studies from the field of education. Lastly, Springer Link and Web of Science were also used, which targets more general fields in cognitive science, including scientific theories of learning, memory, cognition, mathematics, and computer science framework models. Table 1 below lists all of the search strings used and the number of returned articles for each search string. From there, the duplicates were removed, and inclusion criteria were applied.

TABLE I. Search parameters

Database	Search String	Additional options	Returned Articles
ERIC	"intelligent tutoring system" AND ("learning" OR "performance")	Peer-reviewed only box checked. Limited to journal articles.	300
PsycInfo	"intelligent tutoring system" AND "improvement" NOT "emotion" NOT "regulation"	Scholarly (Peer Reviewed) Journals box checked. Academic Journals box checked. Language: English box checked.	40
Springer Link	"intelligent tutoring system" AND ("improvement" OR "learning gain") NOT "emotion" NOT "regulation"	Journal article box checked.	2411
Web of Science	"intelligent tutoring system" AND ("improvement" OR "learning gain") NOT "emotion" NOT "regulation"	Journal article box checked.	2127

C. Data Analysis

Once the final data set was compiled, each of the papers was qualitatively analyzed for multiple characteristics of the study, namely: (1) ITS used, (2) subject domain, (3) learning type, (4) scaffolding strategy, (5) system design, (6) experiment design used, (7) learning outcomes, and (8) target population. Also, metadata such as the date of publication was analyzed for trends.

IV. Results

The final data set is comprised of a total of 22 papers that meet our criteria. The literature revealed that most of the research on ITS have been explorative. The designing of the systems is generally derived from at least one learning theory. In addition, the systems may also utilize some creative licenses such as the personality of the agents, or the presentation of the graphics. Relatively few papers have made thorough experimental assessments of the effectiveness of learning using the tutoring systems with different experimental conditions. Many other experiments were not done focusing specifically on

improvements of learning but on usability or psychological aspects of using the system, such as perceptions of the agents, emotion regulation, and helpfulness of the feedbacks given. Discussions that tie back to the original learning theories, as well as improvements on different modes of learning, are also few and far in between. Appendix A presents an overview of the studies identified. Each of the patterns pertaining to the research questions are discussed in detail in the following sections.

A. Overall Trends

Majority of the ITS target population was college students (n=16). The second most prevalent population was high school and middle school students (n=5). One tutoring system, CIRCSIM-Tutor on physiology, targets professional medical students. The publication trend reveals that the ITS on the domain of physics are the earliest systems that were built. The earlier models tended to use canned feedback expressions and student knowledge model to effectively tackle student misconception. The systems used guided learning with steps through asking questions. However, dialogues with agents was quickly adapted shortly after, within merely a couple of years. This suggests that the importance of agents were always in the minds of the designers for mimicking more realistic human interactions. Indeed, research seem to suggest that the presence of agents did contribute to a larger learning gain. However, there are currently no in-depth analyses on under which conditions should the agents be needed. From the search it seemed that the domains in which the tasks are more explorative, such as database design, computer science, or mathematics that focus exclusively on problem solving, did not incorporate agent dialogues. This could be because the explorative nature of the task would render the agents as a distraction. The secondary reason could be implementing the agents would prove to be difficult as the programming of the agents must be connected to a specific placeholder, which would be difficult for an explorative task.

B. Application Domains

There were a total of 21 ITS that have been tested for learning improvements. Out of these tutoring systems, five were on the domains of computer science, five were on the domain of physics, four were on the domain of mathematics, two are on the domain of biological and physiological sciences, two are on the domain of research methodology, 1 is on the domain of computer literacy, one is on the domain of database design, and the last one is on the domain of urban science (architectural sciences). From the search results, it is suggested that computer science, physics, and mathematics constitute the majority of the ITS that had been designed and built in the last 20 years.

C. Type of Learning

As stated previously, the current study focuses on three types of learning modes: conceptual learning, problem solving, and model building. The ITS for sciences such as physics, biology and research methods generally focused on conceptual learning. The systems implemented conceptual learning via testing students' shallow and deep knowledge on the domains, usually in the form of asking questions regarding specific concepts. The ITS for mathematics and computer science tend to focus on both conceptual learning and problem solving. The problem solving portions are usually quantitative, and present students with scenarios of problems that the students must use their knowledge to solve. The model building learning mode has been lacking in these tutoring systems. Any model building modules with the current tutoring systems tend to be embedded within problem solving and it's hard to clearly separate the two learning modes. This suggests that implementing this particular learning mode has shown to be a challenge.

D. System Features and Scaffolding

As expected, majority of the modern ITS use agents and dialogue moves. However, different dialogue moves and speech implementation were used in different systems. From the search it has been found that 10 ITS use some types of natural language in order to "converse" with the learner, as well as making assessment of the learner's domain knowledge and expertise level. AutoTutor has also recently explored multiple agents which prompts multi-way conversations such as trialogues (i.e. a conversation between a student agent, a teacher agent, and the learner) [14]. Multi-way conversations can simulate vicarious learning, where the learner can learn by the virtue of a conversation between at least two agents without the input of the learner. Such dialogue facilitation is designed to help students overcome any potential frustration or demotivation. It is important to note that within any dialogue-guided ITS, multiple learning theories have been used to make the system "smarter." AutoTutor for example uses step-based expectation misconception scaffolding [18] but also vicarious learning theories [e.g., 21]. Whereas other ITS may use dialogues simply as a means to give feedback but use other approaches such as Bayesian or constraint modeling.

The ITS that do not use dialogues or other types of conversations typically use some form of model building. For example, the Conceptual Helper [22] uses model-tracing in which the system matches the problems with the in-built expert's solution model, and then uses a probabilistic assessment to guide the remediation. Similarly, the CPP-Tutor [23] uses logical modules embedded with feedback loops which will systematically give canned feedback when the logical modules detect any incorrect responses from the learner. ViPS builds student modeling based on detecting student misconceptions and ask the students to correct the misconceptions [24].

Overall, all ITS utilized some form of learner-centered modeling. The overall logic is having accurate assessment of the learner's knowledge level and expertise, as well as aptly identifying the leaner's potential misconceptions of the topics. The feedbacks given in the systems are generally canned—meaning they are pre-written for potential misconceptions based on the previous experience of expert human tutors. This suggests that none of the tutoring systems are capable of adapting to potential new misconceptions that have not been identified from previous student encounters. However, given the breadth of experience and the limitations to the domains that have been experimented on, such designs largely seem to be capable of handling vast majority, if not all, of the potential misconceptions that the systems will encounter.

E. Experiment Design and Learning Outcomes

From the search, vast majority of the experiments that have been conducted used at least one control condition with a pretest and a posttest. A total of 12 papers out of the pool of 22 used

only one experimental condition and one control condition, with a pretest and a posttest. One paper had an experimental design using two control conditions, and other studies used experimental designs that tested against differential levels of experimental conditions. For example, Graesser, Jackson, Mathews, et al. [25] tested the effectiveness of AutoTutor against reading textbooks and doing nothing. They found that reading textbooks have the same effect on learning as doing nothing. In another paper, Rose, Jordan, Ringenberg, et al. [26] tested their Atlas-Andes systems using either a dialogue vs. no dialogue condition and found that dialogues seemed to have contributed a .9 effect size. Currently, no in-depth analyses using qualitative methods have been found in the search. The learning outcomes from these studies range from no evidence of learning impact to an effect size of 1.24. Overall it has been found that the tutoring systems result in significant learning improvements. However, there is currently no evidence that the learning improvements match those of expert human tutors. This suggests that the tutoring systems can only be used as a learning guide, not to replace classroom learning or human tutoring sessions if needed.

V. Discussion and Implications

The goal of the current study is to review the most up-to-date literature on intelligent tutoring systems and their effectiveness for helping students learn in a STEM domain. We focused our search on the following main criteria: mode of learning, learning theory, scaffolding strategy, system design, and learning gains. In addition, we also looked at the specific STEM subject that was utilized in these systems and their target population.

A. Implications for Learning

Findings from this literature review summarize the benefits of integrating ITS as supplements to teaching STEM-related concepts. Overall, we've found that, although there is an abundance of research on ITS, most of the research has been focused on the developmental aspects of the system, notably making the systems more and more intelligent and mimicking the human tutoring sessions. There is considerable research on the effectiveness of learning, and several meta-analyses have been conducted on learning gains. However, our search reveals a need for more fine-grained research on the effectiveness of conceptual learning, problem-solving, and model building, particularly in STEM domains. We emphasize that the investigation of the learning effectiveness of the tutoring systems should be conducted in reference to the utilization of different learning theories and pedagogical methods, thereby shed light on finding the most optimal combinations of these to further the field of learning aid technologies and cyberlearning. In addition, a greater variety of ITS on STEM domains would help a wide population of students achieving in K-12 and higher education and better opportunities for STEM-related careers.

B. Implications for Learning Design

Although most of the ITS were grounded in at least one learning theory, notably the classical social and cognitive constructivist theory of learning, many of them are also designed in a more exploratory fashion, based on intuitions and loose applications. We believe that the system design of the tutoring systems should always begin with a well-established pedagogy in mind, such as model-based inquiry [27], argumentation-driven inquiry pedagogy [28], or self-explanation in multimedia learning [29]. The research papers written for the results of the effectiveness of ITS should also clearly identify the pedagogies used, as well as the core learning theory(ies) being utilized, before describing the scaffolding strategies in its system design and architectural components. This way, it can make it easier for educators for STEM education to establish which of the tutoring systems is a best fit for their students and is compatible and consistent with the classroom and lesson curriculums.

C. Implications for Education Research

From our literature search, we would like to make a case for more qualitative studies that can investigate more in-depth the effectiveness of different learning modes, pedagogy along with its scaffolding strategies, and learning theories. The usage of qualitative methods not only can help investigate learning gains, but it will also help us gain a deeper insight into how individual students learn and the aforementioned factors that can determine the best system design of an ITS in STEM education. It is also important to point out that the findings of the current study are limited and bounded by the search criteria. Thus, many of the qualitative studies may have been excluded by this reason.

VI. Conclusions

ITS have made important contributions to help students gain deep knowledge in STEM education domains. Although many quantitative research studies have investigated the overall learning gains of these tutoring systems, we recommend a focus on qualitative methodology to investigate the core components of pedagogy, learning theories, and modes of learning. Qualitative methods can reach insights that quantitative methodologies currently cannot, such as individual student preferences, progression, and specific needs. With these insights, the system can truly become smarter and become fully adaptive and personalized for individual student learning, which we emphasize is a need that has been lacking in effective STEM instruction and tutoring.

References

[1] NCES. (2013). National Centers for Education Statistics, U.S. Department of Education. Available: https://nces.ed.gov/

[2] J. Roschelle, W. Martin, J. Ahn, and P. Schank, "Cyberlearning community report: The state of cyberlearning and the future of learning with technology," SRI International2017.

[3] A. C. Graesser, "Conversations with AutoTutor help students learn," International Journal of Artificial Intelligence in Education, vol. 26, pp. 124-132, 2016.

[4] A. C. Graesser, N. Dowell, A. J. Hampton, A. M. Lippert, H. Li, and D. W. Shaffer, "Building intelligent conversational tutors and mentors for team collaborative problem solving: Guidance from the 2015 Program for International Student Assessment," in Building Intelligent Tutoring Systems for Teams, ed: Emerald Publishing Limited, 2018.

[5] A. C. Graesser, S. D'Mello, X. Hu, Z. Cai, A. Olney, and B. Morgan, "AutoTutor," in Applied natural language processing: Identification, investigation and resolution, ed: IGI Global, 2012, pp. 169-187.

[6] K. C. Powell and C. J. Kalina, "Cognitive and social constructivism: Developing tools for an effective classroom," Education, vol. 130, 2009.

[7] L. Vygotsky, "Interaction between learning and development," Readings on the development of children, vol. 23, pp. 34-41, 1978.

[8] T. Murray and I. Arroyo, "Toward measuring and maintaining the zone of proximal development in adaptive instructional systems," in International conference on intelligent tutoring systems, 2002, pp. 749-758.

[9] R. Moreno, "Instructional technology: Promise and pitfalls," Technology-based education: Bringing researchers and practitioners together, pp. 1-19, 2005.

[10] B. Lehman, S. D'Mello, A. Strain, C. Mills, M. Gross, A. Dobbins, et al., "Inducing and tracking confusion with contradictions during complex learning," International Journal of Artificial Intelligence in Education, vol. 22, pp. 85-105, 2013.

[11] V. A. Aleven and K. R. Koedinger, "An effective metacognitive strategy: Learning by doing and explaining with a computer - based cognitive tutor," Cognitive science, vol. 26, pp. 147-179, 2002.

[12] M. T. Chi, N. De Leeuw, M.-H. Chiu, and C. LaVancher, "Eliciting self-explanations improves understanding," Cognitive science, vol. 18, pp. 439-477, 1994.

[13] M. T. Chi, S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann, "Learning from human tutoring," Cognitive science, vol. 25, pp. 471-533, 2001.

[14] Z. Cai, S. Feng, W. Baer, and A. Graesser, "Instructional Strategies in Trialogue-based Intelligent Tutoring Systems," Design recommendations for intelligent tutoring systems, vol. 2, pp. 225-235, 2014.

[15] A. Mitrovic, S. Ohlsson, and D. K. Barrow, "The effect of positive feedback in a constraint-based intelligent tutoring system," Computers & Education, vol. 60, pp. 264-272, 2013.

[16] P. Suraweera and A. Mitrovic, "KERMIT: A constraint-based tutor for database modeling," in International Conference on Intelligent Tutoring Systems, 2002, pp. 377-387.

[17] S. Ritter, J. R. Anderson, K. R. Koedinger, and A. Corbett, "Cognitive Tutor: Applied research in mathematics education," Psychonomic bulletin & review, vol. 14, pp. 249-255, 2007.

[18] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," Educational Psychologist, vol. 46, pp. 197-221, 2011.

[19] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu, "Intelligent tutoring systems and learning outcomes: A meta-analysis," Journal of educational psychology, vol. 106, p. 901, 2014.

[20] J. A. Kulik and J. Fletcher, "Effectiveness of intelligent tutoring systems: a meta-analytic review," Review of Educational Research, vol. 86, pp. 42-78, 2016.

[21] M. J. Fryling, C. Johnston, and L. J. Hayes, "Understanding observational learning: An interbehavioral approach," The Analysis of verbal behavior, vol. 27, pp. 191-203, 2011.

[22] P. L. Albacete and K. VanLehn, "Evaluating the effectiveness of a cognitive tutor for fundamental physics concepts," in Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society, 2000, pp. 25-30.

[23] S. S. Abu-Naser, "Evaluating the effectiveness of the CPP-Tutor, an Intelligent Tutoring System for students learning to program in C++," 2009.

[24] L. S. Myneni, N. H. Narayanan, S. Rebello, A. Rouinfar, and S. Pumtambekar, "An interactive and intelligent learning system for physics education," IEEE Transactions on learning technologies, vol. 6, pp. 228-239, 2013.

[25] A. C. Graesser, G. T. Jackson, E. Matthews, H. H. Mitchell, A. Olney, M. Ventura, et al., "Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog," in Proceedings of the Annual Meeting of the Cognitive Science Society, 2003.

[26] C. P. Rosé, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein, "Interactive conceptual tutoring in Atlas-Andes," in Proceedings of AI in education 2001 conference, 2001, pp. 151-153.

[27] M. Windschitl, J. Thompson, and M. Braaten, "Beyond the scientific method: Model - based inquiry as a new paradigm of preference for school science investigations," Science education, vol. 92, pp. 941-967, 2008.

[28] V. Sampson, J. Grooms, and J. P. Walker, "Argument - Driven Inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study," Science Education, vol. 95, pp. 217-257, 2011.

[29] M. Roy and M. T. Chi, "The self-explanation principle in multimedia learning," The Cambridge handbook of multimedia learning, pp. 271-286, 2005.

[30] A. C. Graesser, K. Moreno, J. Marineau, A. Adcock, A. Olney, N. Person, et al., "AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head," in Proceedings of artificial intelligence in education, 2003.

[31] A. M. Olney, S. D'Mello, N. Person, W. Cade, P. Hays, C. Williams, et al., "Guru: A computer tutor that models expert human tutors," in International conference on intelligent tutoring systems, 2012, pp. 256-261.

[32] C. Forsyth, P. Pavlik Jr, A. C. Graesser, Z. Cai, M.-l. Germany, K. Millis, et al., "Learning Gains for Core Concepts in a Serious Game on Scientific Reasoning," International Educational Data Mining Society, 2012.

[33] V. Rus, N. Niraula, and R. Banjade, "DeepTutor: An effective, online intelligent tutoring system that promotes deep learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2015.

[34] P. Nash and D. W. Shaffer, "Mentor modeling: The internalization of modeled professional thinking in an epistemic game," Journal of Computer Assisted Learning, vol. 27, pp. 173-189, 2011.

[35] E. Arnott, P. Hastings, and D. Allbritton, "Research Methods Tutor: Evaluation of a dialogue-based tutoring system in the classroom," Behavior Research Methods, vol. 40, pp. 694-698, 2008.

[36] C. R. Beal, R. Walles, I. Arroyo, and B. P. Woolf, "On-line tutoring for math achievement testing: A controlled evaluation," Journal of Interactive Online Learning, vol. 6, pp. 43-55, 2007.

[37] J. V. Cabalo, B. Ma, and A. Jaciw, "Comparative Effectiveness of Carnegie Learning's" Cognitive Tutor Bridge to Algebra" Curriculum: A Report of a Randomized Experiment in the Maui School District. Research Report," Empirical Education Inc., 2007.

[38] T. C. Chien, M. Yunus, A. Suraya, W. Z. W. Ali, and A. Bakar, "The Effect of an Intelligent Tutoring System (ITS) on Student Achievement in Algebraic Expression," Online Submission, vol. 1, pp. 25-38, 2008.

[39] D. Fossati, B. Di Eugenio, S. Ohlsson, C. W. Brown, L. Chen, and D. G. Cosejo, "I learn from you, you learn from me: How to make iList learn from students," in AIED, 2009, pp. 491-498.

[40] G. Hagerty and S. Smith, "Using the web-based interactive software ALEKS to enhance college algebra," Mathematics & Computer Education, vol. 39, 2005.

[41] Z. Jeremic, J. Jovanovic, and D. Gasevic, "Evaluating an intelligent tutoring system for design patterns: The DEPTHS experience," Journal of Educational Technology & Society, vol. 12, p. 111, 2009.

[42] A. N. Kumar, "Model-based reasoning for domain modeling in a web-based intelligent tutoring system to help students learn to debug c++ programs," in International Conference on Intelligent Tutoring Systems, 2002, pp. 792-801.

[43] H. C. Lane and K. VanLehn, "Teaching the tacit knowledge of programming to noviceswith natural language tutoring," Computer Science Education, vol. 15, pp. 183-201, 2005.

[44] S. Stankov, V. Glavinić, and A. Grubišić, "What is our effect size: Evaluating the educational influence of a web-based intelligent authoring shell," in Proceedings INES 2004/8th International Conference on Intelligent Engineering Systems, 2004, pp. 545-550.

[45] C. W. Woo, M. W. Evens, R. Freedman, M. Glass, L. S. Shim, Y. Zhang, et al., "An intelligent tutoring system that generates a natural language dialogue using dynamic multi-level planning," Artificial intelligence in medicine, vol. 38, pp. 25-46, 2006.

APPENDIX A. OVERVIEW OF THE TWENTY-TWO PAPERS THAT MEET THE SEARCH CRITERIA.

Authors	Domain	Learning Type	System Design	Scaffolding Strategy	Research Question	Environment Used	Target Population	Experiment Design	Learning Outcome
Graesser, et al., 2003 [25]	Physics	Conceptual Learning	Agent and LSA	Natural Language Dialogues Moves	How does learning through AutoTutor compare with reading a textbook or doing nothing?	AutoTutor	College Students	tested against reading textbook and doing nothing	Effect size of .75 against pretest, .61 against control condition, 1.22 against reading textbook; When used pre-post gain scores, AutoTutor>Textbook=Control
Graesser, Moreno, Marineau, Adcock, Olney, & Person, 2003 [30]	Computer Literacy	Conceptual Learning	Agent and LSA	Natural Language Dialogues Moves	Does having a talking head agent help students learn better in an intelligent tutoring system?	AutoTutor	College Students	3(AutoTutor, Read-text, Control)X4 ((Print, Speech, Talking Head, TH+Print) experimental design	Scores on the Replicate previous findings, having a talking head had slightly better benefit with higher proportions of correct responses
Olney, D'Mello, Person, Cade, Hays, Williams, Lehman, & Graesser, 2012 [31]	Biology	Conceptual Learning	Agent and LSA	Preview, Lecture, Summary, Concept Maps	How does learning through the ITS Guru compare with human tutors?	Guru	High schoolers (10th graders)	Educational research/Classroom learning with pre-, post- and delayed posttest. Compared between ITS and human tutor	Effect size of .75 against immediate pretest; human tutor and ITS did not differ significantly; replicated for delayed posttest when compared to classroom only
Forsyth, Pavlik, Graesser, Cai, Germany, Millis, et al. 2012 [32]	Research Methodology	Conceptual Learning	Agent and LSA in Game	critiquing case studies and question generation	Can students learn core concepts of research methodology through a game environment with animated agents?	OperationARIES!	College Students	Case study with pre- and posttest	From .17 (Causal Claims) to .50 (Subject Bias), with a mean of .34 over the 11 core concepts.
Rose, Jordan, Ringenberg, Siler, VanLehn, & Weinstein, 2001 [26]	Physics	Conceptual Learning	LSA	Natural Language Dialogues Moves	What is the learning gain for learning physics with Atlas-Andes?	Atlas-Andes	College Students	experimental condition with or without dialogue	.9 effect size
Rus, Niraula, & Banjade, 2015 [33]	Physics	Conceptual learning and problem solving	Deep dialogue management and natural language understanding components; macro-adaptivity	Self-explanation learning strategies, dialogue, feedback, and learning progressions (LPs)	Does implementing learning progressions better assist student learning and make better predictions regarding student knowledge in an ITS environment?	DeepTutor	High Schoolers	Experimental design with two conditions and pre and posttest for learning gain	Long term learning gain of .43
Nash & Shaffer, 2011 [34]	Urban Science	Conceptual learning and problem solving	epistemic gaming with mentor assessment	Epistemic frame theory	Does collaborative learning in a game environment with a mentor help individual students	Urban Science	Middle School	Experimental design with pre and post interviews	the weighted density of the players' post interview frames was

					achieve knowledge?				significantly greater than that of their pre-interviews (mean pre = 0.1, mean post = 4.4)
Abu-Naser, 2009 [23]	Computer Programming	Problem solving	professional programming integrated development environment	Feedback loop using logical modules	Can students' problem solve programming issues using CPP-Tutor?	CPP-Tutor	College Students	Experiment with control and test; with pre and posttest	10-17% better performance on posttest
Albacete & VanLehn, 2000 [22]	Physics (Mechanics)	Conceptual Learning	model-tracing using probabilistic assessment to guide the remediation	Knowledge base linking	Can Conceptual Helper help student learn mechanics?	Conceptual Helper	College Students	Experiment with control and test; with pre and posttest	.43-.63 effect size
Arnott, Hastings, & Allbritton, 2008 [35]	Research Methodology	Conceptual Learning	Agent and LSA	Natural Language Dialogues Moves	Can Research Methods Tutor help students learn research methodology?	Research Methods Tutor	College Students	Experiment with control and test; with pre and posttest	.75 effect size
Beal, Arroyo, Cohen, & Woolf, 2010 [36]	Mathematics	Problem solving	Agent and LSA	Natural Language Dialogues Moves	Is learning using AnimalWatch better for small group learning or big group learning?	Animal Watch	Middle School	Experiment with pre and posttest; set conditions to varying levels and small vs. big group conditions	significant improvements, especially for low-level students. No effect size was reported
Cabalo, Jaciw, & Vu, 2007 [37]	Mathematics	Problem solving	print and electronic materials,	multiple learning style and step-by-step demonstration of problem	Can CognitiveTutor help student learn mathematics?	CognitiveTutor	College Students	randomized control experiments	No evidence of impact, two experiments show negative effect size
Chien, Yunus, Ali, & Bakar, 2008 [38]	Mathematics	Problem solving	Agent and LSA	Natural Language Dialogues Moves	Can Ms Lindquist help student learn Mathematics?	Ms Lindquist	Middle School	experimental design with two conditions and pre and post interviews	significantly more effective than simply computer assisted learning
Fossati, Eugenio, Brown, & Ohlsson, 2008 [39]	Computer Science	Conceptual Learning	linked lists that can be seen and manipulated	Constraint-Based Modeling-domain knowledge modeled with set of constraints	Can iList help students learn computer science?	iList	College Students	Experimental design with pre and posttest and questionnaire; compared with human tutors	better than control but worse than human tutors
Hagerty & Smith, 2005 [40]	Mathematics	Conceptual learning	adaptive questioning to assess student knowledge	Mastery learning techniques based on knowledge space theory	Can ALEKS help student learn mathematics and contribute to long term retention?	ALEKS	College Students	experimental design with two conditions and pre and post interviews	several assessments show that students using ALEKS improved in posttest and learning retention

Jeremic, Jovanovic, & Gasevic, 2009 [41]	Compute r Science	Conceptual learning	Modules exploration	Content and link level adaptation (adaptive navigation support) using both direct guidance and link removal	Can DEPTHS help student learn computer science?	DEPTH S	College Students	Experiment al design with on test group and two control group	significantly more effective than both control groups
Kumar, 2002 [42]	Compute r Science	Problem solving	feedback and explanation	Model-based reasoning	Can an ITS help student perform better in learning programming?	a C++ tutor (name not given)	College Students	experiment al design with test and control and pre and posttest	2.16 effect size
Lane & VanLehn, 2005 [43]	Compute r Science	Conceptual learning and problem solving	Agent and LSA	Natural Language Dialogues Moves	Can Pro-PELL help student learn computer science?	Pro-PELL	College Students	experiment al design with test and control and pre and posttest	.57 to 2.33 effect size
Myneni & Narayanan, 2012 [24]	Physics	Conceptual learning	student modeling	Misconcepti on detection	Can ViPS help student learn physics?	ViPS	College Students	Experiment al design with three conditions for pre and post	eta-squared of .28 and power of .56
Stankov, Glavinic, & Grubisic, 2004 [44]	Compute r Science	Conceptual Learning	test generation, student rating, student progress and observation	Scaffolding through user domain knowledge	Can DTEx-Sys help student learn computer science?	DTEx-Sys	College Students	Experiment al design with three conditions adding pre and post	.94 effect size
Suraweera & Mitrovic, 2002 [16]	Database design	Problem solving	feedback and multimedia	Constraint-Based Modeling	Can Kermit help students learn Database design?	Kermit	College Students	Experiment al design with test and control and pre and posttest	.63-.66 effect size
Woo, Evens, Freedman, Glass, Shim, Zhang, Zhou, & Michael, 2006 [45]	Physiolo gy	Problem solving	Agent and LSA	Natural Language Dialogues Moves	Can CIRCSIM-Tutor help student learn physiology?	CIRCSI M-Tutor	Medical students	Experiment al design with test and control and pre and posttest	.54 - 1.24 effect size