Education and Information Technologies (2023) 28:9113–9148
https://doi.org/10.1007/s10639-022-11555-x

Examining the applications of intelligent tutoring systems in real educational contexts: A systematic literature review from the social experiment perspective

Huanhuan Wang[1] · Ahmed Tlili[2] · Ronghuai Huang[1] · Zhenyu Cai[3] · Min Li[1] · Zui Cheng[4] · Dong Yang[2] · Mengti Li[5] · Xixian Zhu[1] · Cheng Fei[2]

Received: 28 July 2022 / Accepted: 20 December 2022 / Published online: 7 January 2023

Abstract

Intelligent Tutoring Systems (ITSs) have a great potential to effectively transform teaching and learning. As more efforts have been put on designing and developing ITSs and integrating them within learning and instruction, mixed types of results about the effectiveness of ITS have been reported. Therefore, it is necessary to investigate how ITSs work in real and natural educational contexts and the associated challenges of ITS application and evaluation. Through a systematic literature review method, this study analyzed 40 qualified studies that applied social experiment methods to examine the effectiveness of ITS during 2011–2022. The obtained results highlighted a complicated landscape regarding the effectiveness of ITS in real educational contexts. Specifically, there was an "intelligent" regional gap regarding the distribution of countries where ITS studies using social experiment methods were conducted. Compared to learning performance, relatively less attention was paid to investigating the impact of ITS on non-cognitive factors, process-oriented factors, and social outcomes, calling for more research in this regard. Considering the complexities and challenges existing in real educational fields, there was a lack of scientific rigor in terms of experimental design and data analysis in some of the studies. Based on these findings, suggestions for future study and implications were proposed.

Keywords Intelligent tutoring system · Social experiment · Effectiveness · Challenges · Systematic literature review

✉ Zui Cheng
zc2256@tc.columbia.edu

Extended author information available on the last page of the article

1 Introduction

Artificial Intelligence (AI) in Education is an emerging and booming field (Zawacki-Richter et al., 2019), where one form of its disruptive and transformative application is Intelligent Tutoring Systems (ITSs). ITS refers to computer programs "designed to incorporate techniques from the artificial intelligence (AI) community in order to provide (intelligent) tutors which know what they teach, who they teach, and how to teach it" (Nwana, 1990, p. 252). ITSs can determine the learning path, select and recommend the learning content to students, provide scaffoldings and help engage students in dialogue, and simulate one-to-one tutoring, among others (Zawacki-Richter et al., 2019). They can also provide customized experiences for different students, teachers and tutors (Churi et al., 2022). Thus, ITSs have enormous potentials to support teaching and learning, especially in large-scale distance teaching institutions where human one-to-one tutoring is very difficult (Luckin et al., 2016).

To examine the performance and effectiveness of ITSs, several types of studies have been conducted in the literature. One type of studies focused on evaluating the technical abilities of ITS and answered questions like what ITS can do; other studies focused more on evaluating the effectiveness of ITS as an intervention to improve teaching and learning in a real educational context (Colby, 2017). The second type of studies is particularly important to ITS developers and educational practitioners. From the perspective of ITS developers, ITSs are not just scientists' cool ideas, but also something that should influence the authentic practice of education (Koedinger & Aleven, 2016). Educational practitioners are especially interested in how ITSs can help improve education. Therefore, it is important to conduct field experiments in real educational contexts (i.e., social experiment) to evaluate the effectiveness of ITSs (Corbett et al., 2001).

Social experiment is a research method in which one treatment or more than one alternative treatment are used as interventions into normal social processes and compared (Riecken & Boruch, 1974). This method summarizes the available information about how randomized experiments can be used in planning and evaluating ameliorative programs (Riecken & Boruch, 1974). Applying social experiments to investigate the effectiveness of ITS is crucial, as social experiments can root in real and natural educational environments involving the practice of end users (e.g., students and teachers). Social experiment carefully considers and controls the potential con-founding factors that may affect the observed effectiveness (Riecken & Boruch, 1974). Therefore, it allows discovering causal relationships between the proposed intervention and the effects, and provide reliable evidence to confirm the effectiveness of an intervention (Rolston, 2016).

Despite the volume of literature highlighting the importance of considering social experiment, as a method, to evaluate ITSs, the literature is still fragmented about the practices to do so and the potential challenges. Prior Systematic Literature Reviews (SLRs) on ITS in the literature focused on evaluating the technical performance of ITSs or examining the educational effectiveness of ITSs without considering the critical dimensions of social experiment (e.g., the timespan and

sample size of the conducted studies). No research, to the best of our knowledge, focused on conducting an SLR on ITS from the social experiment perspective. Therefore, to address this gap, this present study aims to systematically review social experiment research investigating the effectiveness of ITSs in real and natural educational contexts.

Different from the previous review studies on ITS, this present study focused on the social experiment perspective when reviewing ITS research and only included and examined studies that focused on the educational effectiveness of ITS in real educational contexts. Conducting such literature review study can help to summarize the outcomes, features and challenges of the ITS research using social experiment methods, hence inform and guide the relevant practices in this context. For educational practitioners, the key take away of such SLRs is that they can provide field evidences on how ITSs work in real educational contexts. For researchers, they can indicate the critical challenges and factors that might influence the observed effectiveness of ITS in education. In addition, a summary of prior studies and factors that influence the success of experimental implementation can guide and inform the practitioners to implement and assess ITSs in the future. Therefore, this study contributes to the literature theoretically and practically. From a theoretical perspective, it enriches the ongoing debate on ITS and social experiment by explaining the current inconsistent results regarding the effectiveness of ITSs on learning and teaching reflected in existing literature. From a practical perspective, this study can support different stakeholders (e.g., ITS researchers and developers, and educational practitioners) learn how to implement ITSs that could effectively work in real and natural educational contexts, keeping in mind several factors (e.g., sample size or interventions).

In the following sections, related literature was reviewed, the detailed systematic review process was reported and findings were analyzed and discussed with their implications.

2 Literature review

2.1 The features of ITSs and their applications in education

ITS can customize instructional activities and strategies based on students' characteristics and needs (Keleş et al., 2009). To provide the desired features, ITSs need to have several components in its system, namely: (1) expert module which contains knowledges for students to learn (Ma et al., 2014); (2) student diagnose module which collects and updates the information about students' knowledge, skills, behaviors, responses, learning styles, etc. (Ma et al., 2014); (3) instructional module which focuses on the strategies and methods of teaching and delivering customized learning content (Carter, 2014); and, (4) user interface which enables the interaction between users and the system (Burns & Capps, 1988).

ITS has been applied in many subject areas to transform teaching and learning. For example, ITSs were used in computer science education to teach students programming skills, followed by medical education and math education

(Mousavinasab et al., 2018). In medical education, ITSs were used to help students learn anatomy, physiology, and diagnosis related knowledge and skills. In mathematics, ITSs were used to facilitate learning numbers, spaces, patterns and structures (Mousavinasab et al., 2018).

To investigate whether ITS has a significant impact on teaching and learning, the effectiveness of ITSs must be evaluated in real and natural educational contexts with proper experimental design, reasonable duration and enough sample size. This evaluation type usually uses field trials or experiments (Koedinger & Aleven, 2016), also known as "social experiment" in social science, as its research method. The major purpose is to evaluate the effectiveness of ITS as an intervention to improve learning and teaching and answer research questions like whether ITS works effectively in a real educational context (Koedinger & Aleven, 2016).

2.2 Social experiment and its features

Social experiment is a research method used in social science, which is defined as a random assignment of participants to two groups to examine the effects caused by social policies (Social experiment, 2008). A social experiment method is a pragmatic trial, with a lot in common with field experiments (Forget, 2019). This method investigates how randomized experiments might be used in planning and evaluating ameliorative social programs (Riecken & Boruch, 1974). In social experiments, one or more treatments are used as interventions and compared (Riecken & Boruch, 1974) to evaluate the effectiveness of the intervention and answer questions like whether the intervention works in the real world (Forget, 2019).

Social experiment has a set of features. Its context is usually set in nonstationary environments in the real world (Fienberg et al., 1985). Since social experiment studies occur in a natural environment, the results can be affected by more "distracting" factors from social, political and economic perspectives. To control the effect of these factors, rigorous experimental design, matching techniques to formulate comparable groups, and advanced analysis technique are often adopted (Rolston, 2016). Participants should ideally be randomly drawn from a specified population and random assignment should ensure that differences in the average behavior of the two groups can be attributed to the treatment. However, in reality, there is less choice beyond basic eligibility; and blinding is usually impossible due to various limitations in the real world (Forget, 2019). The intervention implementation is usually flexible according to the situation in the real world (Forget, 2019). Comparator is essential (Forget, 2019); observations or measurements are used to investigate how some relevant aspects of participants' behaviors differ from those drawn from the same population without treatment. The outcome measures and data collection are directly relevant to stakeholders, such as participants and communities. Social experiments should have evaluative conclusions about the effectiveness of the intervention (Greenberg & Shroder, 2004).

2.3 Examining the effectiveness of ITSs using social experiment methods

Using social experiments to investigate the effectiveness of ITSs is crucial and necessary, since ITSs were not supposed to be effective in principle only, but also as tools that can be integrated in and serve for a full curriculum enhancement (Corbett et al., 2001). Attentions therefore must be paid to the social contexts of schools, training centers or companies where ITSs are used and evaluated (Corbett et al., 2001). When properly implemented, the analysist can ensure that a given intervention has led to a given result (Riecken & Boruch, 1974). Social experiment examines the intervention in real contexts with stakeholders involved. It also considers the potential impact of multiple contextual factors or other con-founding factors and then use rigorous study design, appropriate group matching techniques to formulate comparable control groups, and advanced analytical technique to control the effects of these con-founding factors so that the effect of proposed intervention can be more accurately detected (Rolston, 2016). Such an advantage makes social experiments a strong method of discovering causality (Rolston, 2016).

On the other hand, there exist several crucial challenges related to applying social experiment with ITSs, which can influence the success of experiment implementation, thus affecting the obtained results. These challenges include getting the cooperation of schools to conduct the needed study, handling hardware issues on site, integrating ITS into existing social contexts of schools and instructional practices (Koedinger & Aleven, 2016). However, such a summary of the challenges was drawn from only a few studies, not comprehensively. To support successful application and evaluation of ITS, it is also necessary to comprehensively understand and summarize these challenges reflected in prior related studies.

2.4 Related SLRs focusing on ITS

Several SRLs have been conducted on ITS from different aspects (see Table 1). Some of these reviews focused on ITSs used for domain-specific learning. For example, Neagu et al. (2020) reviewed the studies focused on the efficacy of ITSs in improving

Table 1 The key characteristics of prior literature reviews on ITSs

Dimensions	Characteristics	Examples
Differences	focused on the evaluation of the technical features of ITS	Paladines and Ramirez (2020)
	focused on the evaluation of the effectiveness of ITS in a specific subject	Neagu et al. (2020) Crow et al. (2018) Feng et al. (2021) Alabdulhadi and Faisal (2021) Atun (2020)
	focused on the comprehensive evaluation of ITS	Mousavinasab et al. (2018)
	Focused on the evaluation mediated by ITS	Cuéllar-Rojas et al. (2021)
Limitations	None of the reviews focused on the use of social experiment methods in evaluating ITS effectiveness in real and natural educational contexts across subjects	
Solution	Review studies examining ITS using social experiments and map the overall landscape of the practice of the application and evaluation of ITS in real educational contexts	

on psychomotor training. Crow et al. (2018) reported key information about existing ITSs used for programming education. Feng et al. (2021), and Alabdulhadi and Faisal (2021) reviewed the ITS studies used for supporting STEM-related learning, while Atun (2020) reviewed the ITS studies used to improve reading comprehension.

Other SLRs focused on the evaluation of technological features of ITS. For instance, Paladines and Ramirez (2020) reviewed ITSs incorporating natural dialogue systems. Soofi and Ahmed (2019) reviewed studies that focused on domains, techniques, delivery methods and validation methods of ITS. Cuéllar-Rojas et al. (2021) conducted a systematic review focusing on educational evaluation mediated by ITS. Finally, Mousavinasab et al. (2018) reviewed the overall characteristics, applications, and evaluation method of ITS.

In sum, related SLRs mostly focused on ITS for learning in a specific subject domain, the technical features of ITS, or the overall review of ITS. Yet, none of the aforementioned reviews focused on the use of social experiment methods in evaluating ITS effectiveness in education. The studies examining ITS using social experiments are important and a summary of these studies can map the overall landscape of the practice of the application and evaluation of ITS in real educational contexts and guide future studies. To cover this gap, this study conducts a SLR to synthesize research that adopted social experiment methods to explore the effectiveness of ITS as an intervention in teaching or learning in real and natural educational contexts. Guided by the key features of social experiments pointed out by Riecken and Boruch (1974), Greenberg and Shroder (2004), Forget (2019) and Rolston (2016), this review aims to answer the following research questions (RQs):

- RQ1. What is the trend of ITSs with social experiment research in terms of publication year and the countries where they were applied?
- RQ2. What types of ITS have been utilized and evaluated using social experiment method?
- RQ3. What are the characteristics of ITS research using social experiment method in terms of study contexts, sample size, time span, and study design?
- RQ4. What are the impacts of ITSs through social experiment assessment?
- RQ5. What are the challenges of applying social experiment method to assess the effectiveness of ITS?

3 Methodology

This study followed the recommendation of Kitchenham and Charters (2007) on how to conduct a systematic literature review, which covers three stages, namely: (1) planning the review, which refers to the need for the review and the stated research questions; (2) conducting the review, which refers to the search process of the papers to be included in the review, as well as the data extraction method; and, (3) reporting the review, which describes the way of presenting the results. Each of the three stages are detailed in the next subsequent sections. Additionally, the literature screening followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) proposed by Moher et al. (2015).

Table 2 Search terms

A	B	C	D
1. intelligent*	1. learning	1. system	1. experiment*
2. adaptive	2. instruction	2. software	2. trial
3. customized	3. education	3. application	3. evaluat*
	4. tutoring		4. social experiment
	5. mentoring		

Table 3 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
1. Focused on the evaluation of educational effectiveness of ITS	1. Focused on the evaluation of ability or technical performance of ITS
2. Studies conducted in real learning environments (e.g., classroom, online LMS, training centers, etc.)	2. Studies conducted outside of real learning environment (e.g., technology development lab, etc.)
3. Empirical studies with data or evaluation results	3. Only proposing ITS design solutions or prototypes of ITS without offering evaluation results
4. Articles written in English	4. Articles not written in English
5. Treatment for 8 weeks or more	5. Treatment less than 8 weeks
6. Sample size is equal or greater than 100	6. Sample size is less than 100
7. Paper has been peer reviewed	7. Paper has not been peer reviewed

3.1 Planning the review

A search for studies was conducted in the following databases, which are popular in the field of educational technology, namely: Web of Science, Scopus, IEEE Xplore and ERIC. To deal with the complex topic, the combination of search strings presented in Table 2 were used. Specifically, search terms for ITS were partially adapted from previous reviews (Li & Wong, 2021; Mousavinasab et al., 2018). The asterisk was used to broaden a search. The searching strings were formulated as: ("Intelligent* OR adaptive OR customized) AND (learning OR instruction OR education OR tutoring OR mentoring) AND (system OR software OR application) AND (experiment* OR trial OR evaluat* OR "social experiment").

The obtained papers were then filtered according to the inclusion and exclusion criteria presented in Table 3. To ensure the quality of the obtained results, only peer-reviewed empirical studies published in journals or proceedings were included (Harris et al., 2014). The time frame was set as 2011- 2022, as 2011 was considered the year of where AI assisted people was booming. For instance, IBM's Watson defeated television games and Apple's Seri was released. Thus, it would be important to see how this impacted ITSs which integrate AI to assist teaching and learning. Besides, following the best evidence criterion for ensuring a good external validity proposed by Slavin (1986), this present study only

included experiments that have a time span of 8 weeks and more. Regarding the sample size, we followed Guha (2008)'s suggestion about social experiments and included only studies with a sample size of 100 participants and more. The final search was conducted on October 14th, 2022, which led to finding 21,294 studies from the specified databases and 28 studies identified by going through the references of the obtained articles.

Finally, two authors analyzed the retrieved papers by titles, abstracts, and, if necessary, by full text, based on the pre-defined inclusion and exclusion criteria (see Table 3). Figure 1 presents the flow diagram of the study selection process. At the end of this process, 40 studies were identified as being relevant to the purpose of this present SLR.

3.2 Conducting the review

This stage includes the data extraction process. A coding schema, as shown in Table 4, was developed based on the major components of social experiments indicated in the literature review section to answer the aforementioned research questions. To reduce the opportunity for bias, an electronic data

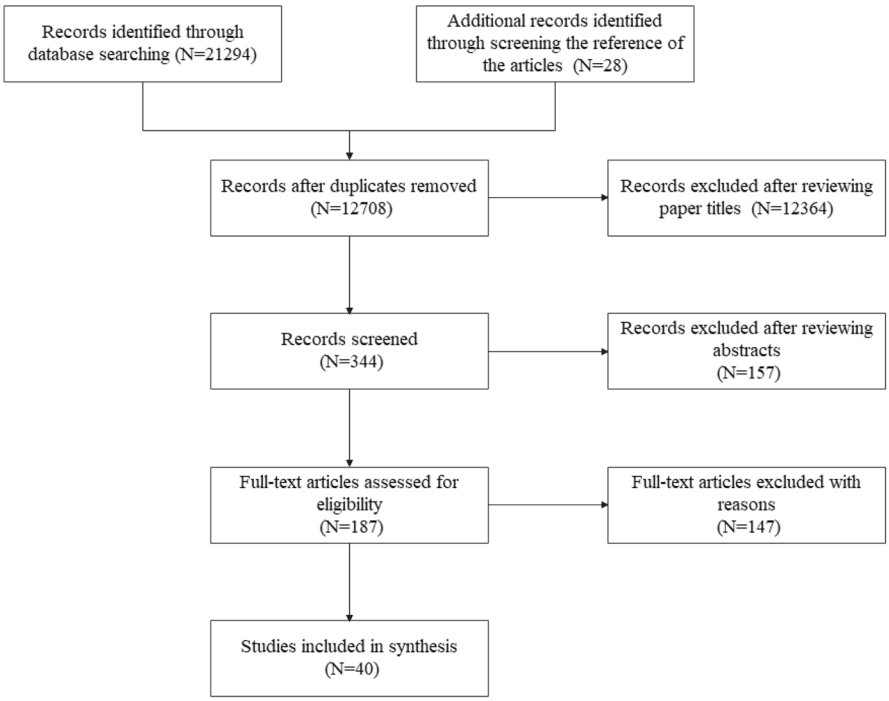


Fig. 1 The studies selection process

Table 4 Description of the coding scheme

Level 1 code	Level 2 codes	Examples
ITS Intervention	ITS name	Auto Tutor
	Major features of ITS	Providing personalized feedback
Study context	Country	China
	Education-level	Higher education
	Subject area	Math
Method	Experiment design	Random Control Trial
	Assignment method	Random assignment
Criteria for benchmarking		Control group where human tutors were used
Results	Outcome variable	Learning performance
	Effects found	ITS can significantly improve learning performance
Challenges	Pedagogical challenges	Consider students' individual difference
	Technological challenges	Equipment is not sufficient
	Methodological challenge	Limited data access

extraction form based on the coding scheme was designed (Kitchenham & Charters, 2007).

3.3 Reporting the review

In this stage, the extracted data, based on the coding scheme, were compared and discussed to answer the research questions.

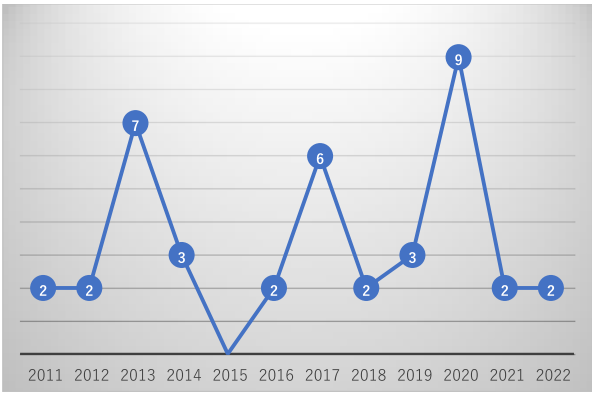


Fig. 2 Frequency of publications on ITS studies applied social experiment method

4 Findings

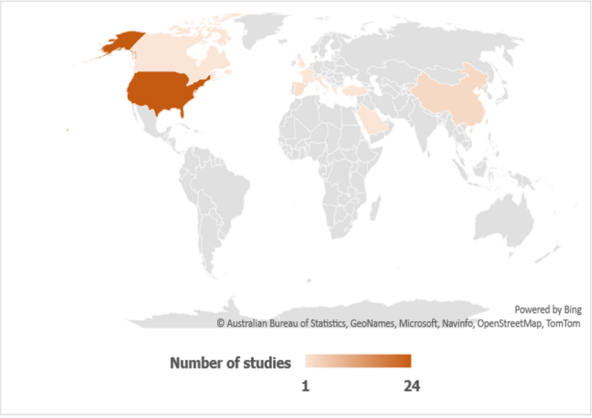
4.1 RQ1. What is the trend of ITSs with social experiment research in terms of publication year and the countries where they were applied?

A total of 40 articles (see Appendix) were finally reviewed and coded based on the coding schema. Figure 2 shows that there were several peaks in 2013, 2017 and 2020 in terms of the number of published studies on ITS with social experiment method. Specifically, 2020 was the year with the highest number of publication (9 studies).

Studies that examined ITS using social experiment methods in the set time span (length ≥ 8 weeks) and sample size (N ≥ 100) have been carried out in several countries, as shown in Fig. 3. Specifically, 60% (n = 24) of these experiments were carried out in the USA, followed by the Netherlands (10%, n = 4) and China (8%, n = 3).

4.2 RQ2. What types of ITS have been utilized and evaluated using social experiment method?

As described in Table 5, five categories of ITSs have emerged which are presented from the largest to the smallest in terms of the number of studies involved: (1) recommendation and tutoring (13, 32.5%), (2) personalized support (12, 25%), (3) exercise and assessment (7, 17.5%), (4) personalization (6, 15%), (5) adaptive conversation (3, 7.5%), and (6) game-based learning (1, 2.5%). Specifically, among all the emerged individual ITSs, Cognitive Tutors were the most frequently applied and evaluated ITSs (7, 17.5%). In addition, two other ITSs which may suggest the trends of ITS development are worth mentioning. First, 100 Nano tutors were embedded in the video lesson to guide students' understanding of narrowly defined skills (Goel & Joyner, 2017). Second, complicated tutors integrating multiple tutors have been invented and used. For



Country/Region	Number of studies
USA	24
Netherlands	4
China	3
Spain	2
Canada	1
Turkey	1
Greece	1
France	1
Saudi Arabia	1
United Kingdom	1
Italy	1

Fig. 3 Regional distribution of the reviewed ITS studies

Table 5 Categories of the applied and evaluated ITSs

Category	ITS name	Main features	Associated studies	N (%)
Recommendation and tutoring	1. Cognitive Tutor	A type of ITSs which provide personalized problems and step-by-step instructions based on cognitive models of students it created	Bernacki and Walkington (2018); Lee et al. (2013); Butcher and Aleven (2013); Roll et al. (2011); Long and Aleven (2017); Fang & Guo (2013); Pane et al. (2014)	13 (32.5%)
	2. ITSS	A web-based ITS integrated with the structure strategy to improve reading comprehension	Wijekumar et al., (2012, 2013, 2014, 2020)	
	3. Dragoon 2	A step-based ITS, which uses example-tracing, an explicit pedagogical policy, and an open learner model, can teach students model construction for dynamic systems	Wetzel et al. (2017)	
	4. PCAR	A personalized context-aware recommendation learning system, which can help students learn English through mobile devices	Yao (2017)	

Table 5 (continued)

Category	ITS name	Main features	Associated studies	N (%)
Personalized support	5. Lexue "100"	A web-based ITS, which can support math-teaching and learning through personalized practice, assessment and guidance	Zhang & Jia (2017)	10 (25%)
	6. Mousework	A web-based ITS, providing students who study at home personalized exercises and feedback for multiple subjects	Bartelet et al. (2016)	
	7. Living Letters	An online ITS, which offers personalized instruction, hints, and corrections to focus students on solving target problems	Kegel and Bus (2012)	
	8. SARA	An ITS which can provide automated and personalized feedback to students	Mousavi et al. (2021)	
	9. FB-TS	An ITS developed based on Fuzzy logic and Bayesian network, which can provide adaptive support to students	Eryilmaz and Adabashi (2020)	
	10. Nanotutors	ITSs, highly focused intelligent agents used in the video lessons and can help students understand AI concepts and skills	Goel and Joyner (2017)	
	11. DME	An ITS which can provide combined inner and outer loop feedback on students' social science learning process and performance	Tacoma et al. (2020)	
	12. ELAN	An adaptive learning game that supports literacy acquisition through teaching and training phonics	Watkins et al. (2020)	
	13. Reading Plus	A web-based learning program with adaptive assessment, which can scaffold silent reading-related learning	Spichtig et al. (2019)	
	14. Other ITSs without a name	Adaptive e-learning system based on situational awareness and remodeled teaching; It can provide adaptive feedback to students	Capone et al. (2022)	

Table 5 (continued)

Category	ITS name	Main features	Associated studies	N (%)
Exercise and Assessment	15. ALEKS	A web-based ITS, which can adaptively select the next skill for students to work on	Craig et al. (2013); Cung et al. (2019); Huang et al. (2013); Hickey et al. (2020)	7 (17.5%)
	16. ASSISTments	A web-based ITS, providing homework intervention in terms of adaptive assessment and feedback	Feng et al. (2014); Jiang et al. (2020)	
	17. Dutch Education system	Adaptive practice program that provides individualized sequences of exercises which can support students to practice the learned content	Klaveren et al. (2017)	
Personalization	18. CTAT	A model tracing ITS, which can trace the interaction of students with the system and recommend the content to learn in the next steps	Treceño-Fernández et al. (2020)	6 (15%)
	19. HINTS	Hypergraph Intelligent Tutoring System (HINTS) can provide personalized learning path	del Olmo-Muñozet al. (2022)	
	20. Other ITSs without names	Guide students through a series of activities to develop their understanding of the selected words	Baker et al. (2020);	
		Provide personalized materials based on its adaptive agent	Chang et al. (2016)	
		Provide adaptive materials and activities to students	Troussas et al. (2021);	
		Provide adaptive materials and activities to students	Alshammari and Qtaish (2019)	
Adaptative conversation	21. SKOPE-IT	A hybrid ITS combining simple ITSs Auto Tutor with ALEKS, which can talk to students to offer learning support	Nye et al. (2018)	3 (7.5%)
	22. My Science Tutor	A new generation of ITS, which facilitates learning through natural spoken dialogs with a virtual tutor in multimedia activities	Ward et al. (2011); Ward et al. (2013)	
Game-based learning	23. GraphoGame Rime	An ITS, which can support language learning (e.g., phonics) through game-based learning	Ahmed et al. (2020)	1 (2.5%)

example, SKOPE-IT combined ITSs Auto Tutor and ALEKS to process more complex tasks (Nye et al., 2018).

4.3 RQ3. What are the characteristics of ITS research using social experiment method in terms of study contexts, sample size, time span, study design and benchmarking used for evaluation?

Table 6 shows that the reviewed studies have been conducted across most of the educational levels, including kindergarten, primary schools, secondary schools, higher education, and adult learning. Specifically, secondary education is where most of the ITS studies (16, 40%) have been conducted, followed by higher education (14, 35%) and primary education (13, 32.5%). However, Adult learning (1, 2.5%) and Kindergarten education (1, 2.5%) were the least investigated using social experiment, calling for further research in this context. Finally, 5% of these studies were implemented across several educational levels. For example, Zhang and Jia (2017) implemented the experiments in both primary schools and secondary schools. Wetzel et al. (2017) conducted their experiments in secondary schools and higher education.

Additionally, ITSs have been applied to support learning in multiple subject areas, such as language, math, science, computer science, medicine, history, economics, geometry, and engineering. Math (33%), language (24%), and science (17%) were the primary subject areas where ITSs were applied to support learning and teaching. In contrast, engineering (5%), history (2%), and economic (2%) were the areas where relatively less studies were conducted (see Fig. 4).

Table 7 shows that 62.5% of the reviewed studies had the number of participants ranging from 100–500. It is noticeable that 7 studies (17.5%) had more than 1000 individual participants. It was also found that 5 studies (Pane et al., 2014; Wijekumar et al., 2013; Wijekumar et al., 2014; Wijekumar et al., 2020; Zhang & Jia, 2017) conducted large-scale experiments and involved the whole school as participants.

According to Table 8, the time span of ITS experiments varied from 8 weeks to 5 years. It is observed that for 55% of the studies, their time span is more than one year, while about 45% of ITS social experiments have the time span less than one year.

For the characteristics of experimental design (Table 9), 22.5% of the studies did not apply random assignment. In addition, five types of experimental design were used, namely (1) Random Control Trial; (2) Quasi-experiment; (3) Natural experiment; (4) Randomized Alternative-Treatment Design; and, (5) Longitudinal Study. Quasi-experiment (45%) was the most frequently used experimental design, followed by Random Control Trial (37.5%). In addition, natural experiment (5%), longitudinal study (10%), and Randomized Alternative-Treatment Design (2.5%) were rarely used. Finally, only 13.5% of the studies (n = 6) used the matching technique to formulate comparable control groups (Cung et al., 2019; Hickey et al., 2020; Mousavi et al., 2021; Pane et al. (2014); Spichtig et al., 2019; Troussas et al., 2021).

To evaluate the effectiveness of ITSs, the included studies used different benchmarks for comparison, which can be classified into three types (see Table 10). First, business-as-usual was used as a benchmark for comparing with experiment groups where ITSs were used. This category included the experiment conditions where there

Table 6 Educational levels where the studies were conducted

Level	N of pub. (Percentage)	Reviewed studies
Secondary Schools	16 (40%)	Klaveren et al. (2017); Bartelet et al. (2016); Feng et al. (2014); Long and Aleven (2017); Jiang et al. (2020); Bernacki and Walkington (2018); Butcher & Aleven (2013); Roll et al. (2011); Lee et al. (2013); Wetzel et al. (2017); Huang et al. (2013); Zhang & Jia (2017); Pane et al. (2014)
Higher education	14 (35%)	Mousavi et al. (2021); Nye et al. (2018); Eryilmaz and Adabashi (2020); Fang & Guo (2013); Hickey et al. (2020); Goel and Joyner (2017); Wetzel et al. (2017); Cung et al. (2019); Troussas et al. (2021); Tacoma et al. (2020); Yao (2017); Chang et al. (2016); Alshammari and Qtaish (2019); Capone et al. (2022);
Primary Schools	13 (32.5%)	Baker et al. (2020); Wijekumar et al. (2012); Wijekumar et al. (2020); Wijekumar et al. (2014); Wijekumar et al. (2013); Craig et al. (2013); Ward et al. (2011); Ward et al. (2013); Spichtig et al. (2019); Zhang & Jia (2017); Watkins et al. (2020); Ahmed et al. (2020); del Olmo-Muñozde et al. (2022)
Adult learning	1 (2.5%)	Treceno-Fernandez et al. (2020)
Kindergarten	1 (2.5%)	Kegel and Bus (2012)

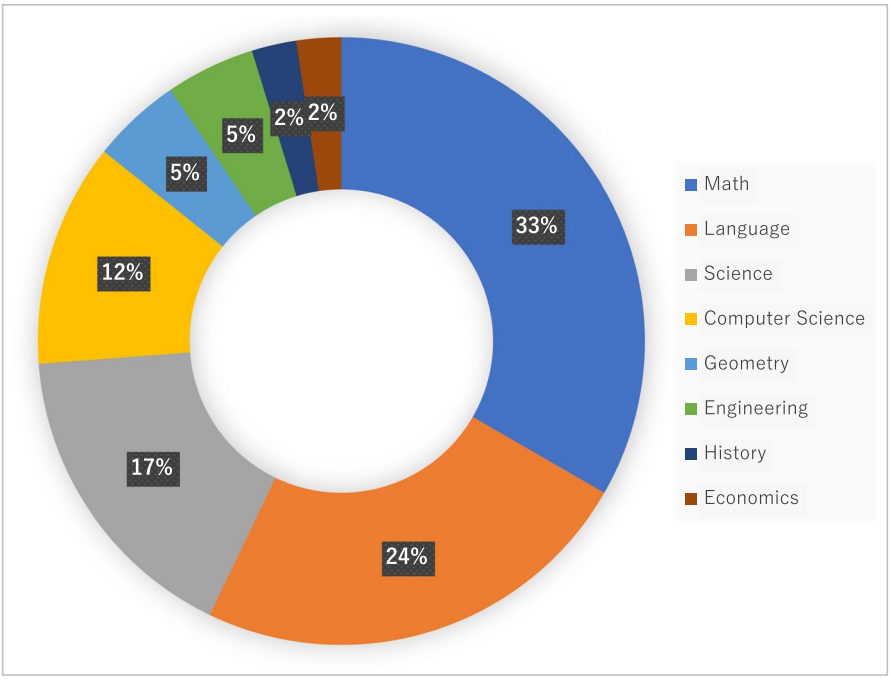


Fig. 4 Distribution of subjects where ITS studies were conducted

Table 7 Sample size of the review ITS studies

Level	Scale	Publications	N (%)
Students	100–500	Craig et al.(2013); Jiang et al. (2020); Kegel and Bus (2012); Bartelet et al. (2016); Huang et al. (2013); Nye et al. (2018); Eryilmaz and Adabashi (2020); Bernacki and Walkington (2018); Ward et al. (2011); Fang and Guo (2013); Treceno-Fernandez et al. (2020); Lee et al. (2013); Wijekumar et al. (2012); Baker et al. (2020); Butcher & Aleven (2013); Wetzel et al. (2017); Roll et al. (2011); Long and Aleven (2017); Troussas et al. (2021); Alshammari and Qtaish (2019); Yao (2017); Spichtig et al. (2019); Chang et al. (2016); Capone et al. (2022); del Olmo-Muñoz et al. (2022)	25 (62.5%)
	500–1000	Watkins et al. (2020); Ahmed et al. (2020); Tacoma et al. (2020)	3 (7.5%)
	1000–3000	Cung et al. (2019); Mousavi et al. (2021); Hickey et al. (2020); Feng et al. (2014); Goel and Joyner (2017); Ward et al. (2013); Klavern et al. (2017)	7 (17.5%)
Schools	19	Zhang & Jia (2017)	1 (2.5%)
	45	Wijekumar et al. (2014); Wijekumar et al. (2020); Wijekumar et al. (2013)	3 (7.5%)
	147	Pane et al. (2014)	1 (2.5%)

Table 8 Time span of the conducted ITS experiments

Time span	Associated studies	Numbers (%)
8 weeks ≤ TS < 1 year	Craig et al. (2013); Kegel and Bus (2012); Bartelet et al. (2016); Huang et al. (2013); Nye et al. (2018); Zhang & Jia (2017); Fang & Guo (2013); Wijekumar et al., (2012, 2014); Baker et al. (2020); Butcher and Aleven (2013); Roll et al. (2011); Troussas et al. (2021); Alshammari and Qtaish (2019); Ahmed et al. (2020); Yao (2017); Spichtig et al. (2019); del Olmo-Muñoz et al. (2022)	18 (45%)
1 year ≤ TS < 2 years	Jiang et al. (2020); Eryilmaz and Adabashi (2020); Bernacki & Walkington (2018); Ward et al. (2011); Lee et al. (2013); Wijekumar et al., (2013, 2020); Watkins et al. (2020); Klaveren et al. (2017); Eryilmaz and Adabashi (2020); Long & Aleven (2017); Ward et al. (2013)	12 (30%)
2 years ≤ TS < 3 years	Cung et al. (2019); Mousavi et al. (2021); Treceno-Fernandez et al. (2020); Feng et al. (2014); Wetzel et al. (2017); Cung et al. (2019); Pane et al. (2014)	7 (17.5%)
3 years ≤ TS < 4 years	Capone et al. (2022)	1 (2.5%)
years 4 ≤ TS < 5 years	Goel and Joyner (2017);	1 (2.5%)
TS ≥ 5 years	Hickey et al. (2020)	1 (2.5%)

Table 9 Study design and assignment method used in the reviewed ITS experiments

Research Design	Random Assignment	Study examples	N (%)
Quasi- experiment	Individual level	Long and Aleven (2017); Huang et al. (2013); Mousavi et al. (2021)*; Nye et al. (2018); Eryilmaz and Adabashi (2020); Bernacki and Walkington (2018); Ward et al. (2011); Ward et al. (2013); Alshammari and Qtaish (2019); Yao (2017); Tacoma et al. (2020); Capone et al. (2022)	18 (45%)
	Group level	del Olmo-Muñoz et al. (2022)	
	None	Zhang et al. (2017); Hickey et al. (2020) *; Cung et al. (2019) *; Troussas et al. (2021) *; Chang (2016)	
Random Control Trial	Group level	Wijekumar et al. (2014); Wijekumar et al. (2012); Wijekumar et al. (2020); Wijekumar et al. (2013); Baker et al. (2020); Feng et al. (2014); Kegel and Bus (2012); Bartelet et al. (2016); Lee et al. (2013); Pane et al. (2014)*	15 (37.5%)
	Individual level	Jiang et al. (2020); Watkins et al. (2020); Klaveren et al. (2017); Ahmed et al. (2020); Spichtig et al. (2019) *	
Longitudinal study	Individual level	Butcher & Aleven (2013); Roll et al. (2011)	4 (10%)
	None	Wetzel et al. (2017); Goel and Joyner (2017)	
Natural experiment	None	Fang and Guo (2013); Treceno-Fernandez et al. (2020)	2 (5%)
Randomized Alternative-Treatment Design	Conditions randomly alternated	Craig et al. (2013)	1 (2.5%)

Note. *Used matching techniques to select the subjects and formulate comparable control group

Table 10 The criteria utilized in benchmarking and evaluating ITS

The benchmarking utilized in evaluation	Related studies	Quantity and Percentage
Business as usual without ITS	Craig et al. (2013); Bartelet et al. (2016); Huang et al. (2013); Eryilmaz and Adabashi (2020); Ward et al. (2011); Zhang et al. (2017); Treceno-Fernandez et al. (2020); Hickey et al. (2020); Wijekumar et al. (2014); Feng et al. (2014); Wijekumar et al. (2012); Baker et al. (2020); Wijekumar et al. (2020); Butcher & Aleven (2013); Wetzel et al. (2017); Goel and Joyner (2017); Wetzel et al. (2017); Ward et al. (2013); Watkins et al. (2020); Ahmed et al. (2020); Yao (2017); Spichtig et al. (2019); Chang et al. (2016); Pane et al. (2014); Kegel and Bus (2012)	25 (62.5%)
ITS providing alternative treatments	Jiang et al. (2020); Mousavi et al. (2021); Nye et al. (2018); Bernacki and Walkington (2018); Fang and Guo (2013); Lee et al. (2013); Roll et al. (2011); Long & Aleven (2017); Troussas et al. (2021); Klaveren et al. (2017); Alshammari and Qtaish (2019); Tacoma et al. (2020); del Olmo-Muñozet al. (2022)	13 (32.5%)
Blended mode with ITS and other forms of instruction	Cung et al. (2019); Capone et al. (2022)	2 (5%)

were human tutors used or there was no additional tutoring provided. This is the largest group among the included studies (62.5%, n=25). This result indicated that most of ITS-related social experiment studies considered ITS as a whole component when comparing it with the condition where there were no ITSs used. Due to the potential Blackbox effect caused by viewing ITS as whole, it is difficult to understand which sub-components of ITS were less effective. Second, ITS carrying alternative treatments (32.5%, n=13). For example, ITS that enables personalization based on other student characteristics, providing no personalized content or feedback. This type of evaluation benchmarks is ITSs with alternative features, which can help overcome the shortage of Blackbox effect and help examine which part(s) of an ITS really works. The last category (5%, n=2) is the blended mode which combined the application of ITS and other form of instruction, such as human tutor, which is used as a benchmark. This method directs researchers to compare the pure machine-enable intelligence and the hybrid intelligence based on the blended mode of ITS and human tutors.

4.4 RQ4. What are the impacts of ITSs through social experiment assessment?

As shown in Table 11, learning performance was the most investigated outcome (36, 90%) to measure the effectiveness of ITS, followed by students' help-seeking

Table 11 Outcomes measured in the reviewed studies

Outcomes	N (%)	Studies
Learning performance	36 (90%)	Kegel and Bus (2012); Eryilmaz and Adabashi (2020); Wetzel et al. (2017); Ward et al. (2011); Zhang & Jia (2017); Chang et al. (2016); Ahmed et al. (2020); Eryilmaz and Adabashi (2020); Bernacki and Walkington (2018); Cung et al. (2019); Mousavi et al. (2021); Troussas et al. (2021); Wijekumar et al., (2012, 2013, 2014, 2020); Long and Aleven (2017); Watkins et al. (2020); Alshammari and Qtaish (2019); Baker et al.(2020); Huang et al. (2013); Craig et al. (2013); Hickey et al. (2020); Fang & Guo (2013); Klaveren et al. (2017); Roll et al. (2011); Jiang et al. (2020); Lee et al. (2013); Nye et al. (2018); Tacoma et al. (2020); Ward et al. (2013); Bartelet et al. (2016); Kegel and Bus (2012); Butcher and Aleven (2013); Pane et al. (2014); del Olmo Muñoz et al. (2022)
Help seeking	4(10%)	Roll et al. (2011); Craig et al. (2013); Lee et al. (2013); Jiang et al. (2020)
Engagement	3 (7.5%)	Bartelet et al. (2016); Craig et al. (2013); Capone et al. (2022)
Attitude	1(2.5%)	Pane et al. (2014)
Confidence	1(2.5%)	Pane et al. (2014)
Interest	2(5%)	Bernacki and Walkington (2018); Chang et al. (2016)
Teachers' perception	5(12.5%)	Baker et al. (2020); Ward et al., (2011, 2013); Craig et al. (2013); Feng et al. (2014)

(4,10%), engagement (3, 7.5%) and interest (2, 5%), and teachers' perceptions (5, 12.5%). The impact of ITS on each of the aforementioned outcomes is discussed in the following subsequent sections.

4.4.1 The impact of ITS on learning performance

There are mixed types of results regarding the overall impact of ITSs on learning performance (see Table 12). 62.5% (25) of the studies reported positive effects. Specifically, researchers reported a set of situations where the positive effects of ITS on learning performance were identified. These situations included ITS explaining how to proceed in learning and why correctness (Kegel & Bus, 2012), using novel teaching strategy (Troussas et al., 2021; Wijekumar et al., 2012, 2013, 2014, 2020), having an open learner model (Long & Aleven, 2017), being designed based on cognitive science (Watkins et al., 2020), providing adaptation according to the combination of learning style and knowledge level (Alshammari & Qtaish, 2019), conducting data analysis while controlling confounding factors (Baker et al., 2020; Wijekumar et al., 2012, 2013, 2014); and, targeting students with specific features, such as white male students (Huang et al., 2013), low and middle level achieving students (Bartelet et al., 2016), being blended with human instructor-led instruction (Pane et al. 2014), among others.

Table 12 The effects of ITS on learning performance and associated studies

Findings	Quantity	Reviewed studies
Positive effects	25 (62.5%)	Kegel and Bus (2012); Eryilmaz and Adabashi (2020); Wetzel et al. (2017); Ward et al. (2011); Jiang and Jia (2019); Chang et al. (2016); Ahmed et al. (2020); Eryilmaz and Adabashi (2020); Bernacki and Walkington (2018); Cung et al. (2019); Mousavi et al. (2021); Ward et al. (2013); Troussas et al. (2021); Wijekumar et al., (2012, 2013, 2014, 2020); Long and Aleven (2017); Watkins et al. (2020); Alshammari and Qtaish (2019); Baker et al. (2020); Huang et al. (2013); Bartelet et al. (2016); Pane et al. (2014); de Olmo-Munoz et al. (2022)
No effects	14 (35%)	Craig et al. (2013); Ward et al. (2011); Hickey et al. (2020); Zhang and Jia (2017); Fang and Guo (2013); Klaveren et al. (2017); Butcher and Aleven (2013); Roll et al. (2011); Jiang et al. (2020); Lee et al. (2013); Nye et al. (2018); Tacoma et al. (2020); Ward et al. (2013); Wijekumar et al. (2020)
Negative effects	5 (12.5%)	Bartelet et al. (2016); Roll et al. (2011); Jiang et al. (2020); Kegel and Bus (2012); Butcher and Aleven (2013)

37% of the studies reported no significant effect on learning performance. Specifically, ITSs that provide: multiple templates of problem formats (Jiang et al., 2020), verbal explanations (Lee et al., 2013), tutoring-enhanced interactive solutions (Nye et al., 2018), a combination of outer loop feedback and inner loop feedback (Tacoma et al., 2020), character animation technology (Ward et al., 2013), and communication via spoken dialog and analysis while controlling the effects of covariates (Wijekumar et al., 2020) did not have any significant impact on learning performance.

12.5% of the studies reported negative effects caused by ITS. For example, ITS was found to have a negative effect on learning growth for higher achieving students (Bartelet et al., 2016), increased the error rates related to glossary learning (Roll et al., 2011), and ITS with multiple templates of problem format reduced student efficiency (Jiang et al., 2020). Researchers also warned that ITS can have potential negative effects on students' problem solving when it does not explain how to proceed during learning (Kegel & Bus, 2012) and ITS's instructional scaffolds can reduce students' active processing (Butcher & Aleven, 2013).

4.4.2 The impact of ITS on students' help-seeking, engagement, learning interest, attitude, and confidence

As shown in Table 13, 10% of the studies reported the effect of ITSs on help seeking. Two of them reported positive effect and two of them indicated no effect. Specifically, it is reported that ITS can improve students' help-seeking skills (Roll et al., 2011), and reduce the assistance that students need from teachers (Craig et al., 2013). In contrast, Lee et al. (2013) and Jiang et al. (2020) indicated that ITS did not influence students' hint use or requests. Thus, mixed effects of ITS on help-seeking were found.

Regarding the effects of ITSs on learning engagement**,** there is also a mixed type of results (see Table 13). It was found that students with lower level of skills spent more time on practice tasks with ITS (Bartelet et al., 2016) and increased their situational awareness (Capone et al., 2022). However, Craig et al. (2013) found that there was no significant influence of ITS on learning task involvement.

Table 13 ITS's impact on help-seeking, engagement, and learning interests

Outcome	Positive effect	No effect
Help-seeking	Roll et al. (2011) Craig et al. (2013)	Lee et al. (2013) Jiang et al. (2020)
Engagement	Bartelet et al. (2016); Capone et al. (2022)	Craig et al. (2013)
Learning interest	Bernacki and Walkington (2018) Chang et al. (2016)	None
Attitude	None	Pane et al. (2014)
Confidence	None	Pane et al. (2014)

For learning interest, Bernacki and Walkington (2018) and Chang et al. (2016) both reported that ITS improved students' learning interests (see Table 13). For attitude and confidence in math learning, Pane et al. (2014) reported that ITS did not have any significant effect.

4.4.3 Teachers' perceptions on adopting ITS in teaching and learning

Through the use of ITSs, teachers gained better perceptions of their work, including positive perceptions of student experiences with ITSs (Baker et al., 2020). Teachers felt that students were more enthused and engaged in learning (Feng et al., 2014; Ward et al., 2011, 2013). They also perceived that ITSs reduced their workload (Craig et al., 2013; Feng et al., 2014). With more time saved, teachers focused more on problematic areas identified by the learning reports generated by ITSs, and their work focus shifted from checking the correctness of each problem to explaining and elaborating on the mistakes that students did (Feng et al., 2014).

4.5 RQ5. What are the challenges of applying social experiment method to assess the effectiveness of ITS?

It is reported that the central challenge lies in improving the effectiveness of ITS on learning (Zhang & Jia, 2017). Such challenge may stem from a set of other associated challenges reported by several studies.

First, students' limited task involvement. This limited task involvement has different forms, such as low completion rate in the assignments (Jiang et al., 2020). Particularly for young students, learning with ITS involves regulatory skills that might be too demanding (Kegel & Bus, 2012). Cung et al. (2019), Nye et al. (2018), Roll et al. (2011) and Wijekumar et al. (2012) also mentioned high attrition issues, such as participants' withdrawal or absenteeism. Moreover, insufficient involvement due to small sample size issues is noticed. Huang et al. (2013) and Bernacki and Walkington (2018) also reported small sample size issues, which may cause sample bias and therefore favor the control group (Craig et al., 2013), making it difficult to detect the effects of treatments (Bartelet et al., 2016) or achieve a good generalizability of the conclusions (del Olmo-Muñoz et al., 2022).

Second, handling students' individual differences. students' individual differences can lead to the moderate effect size of the proposed ITS intervention (Kegel & Bus, 2012) or ceiling effect (Bartelet et al., 2016) that limits the ability to detect potential significantly positive effect of ITS interventions. Students' background and personal characteristics can vary from a person to another and from a semester to another (Bernacki & Walkington, 2018; Cung et al., 2019). Thus, how to deal with these individual indifferences can be challenging.

Third, limited resources and competencies. For example, it is reported that there was a lack of computer labs, computers, electricity outages (Wijekumar et al., 2013), high quality video equipment (Roll et al., 2011) or learning

systems (Mousavi et al., 2021). In addition, there can be limited data access (Hickey et al., 2020; Butcher & Aleven, 2013; Jiang et al., 2020), and no sufficient resources for one-on-one tutoring (Ward et al., 2011), applying randomized assignment (Treceño-Fernández et al., 2020; Hickey et al., 2020), and keeping intervention dosage (Nye et al., 2018), grading scheme, and the number of benchmarks (Cung et al., 2019) consistent during the study. Additionally, in some studies, the participants did not have the necessary skills (e.g., keyboarding, Baker et al., 2020) to manage and use ITSs. A lack of available time for an experiment is also an issue (Bartelet et al., 2016; Goel & Joyner, 2017; Wetzel et al., 2017) that hinders the conducted studies from having long-term evidence (Kegel & Bus, 2012).

Fourth, methodology-related challenges. It is reported that social experiments, which can have black box effect, made the researchers cannot separate and measure the specific effects of different ITS components (Long & Aleven, 2017). In addition, social experiment requires a long-time span to capture the focused effect. Thus, as time passed by, how to address implementation dip and maintaining implementation fidelity can be challenging (Cung et al., 2019). In addition, there remains challenges in balancing the cost of designing and developing ITS and the benefits it brings (Bernacki & Walkington, 2018), as well as balancing the ability of ITSs for encouraging student engagement and providing optimal challenge levels (Nye et al., 2018).

Finally, for the study conducted recently, such as Capone et al. (2022), challenges like how instructors quickly adapted to this ITS-mediated remote teaching mode in a short time due the COVID-19 pandemic were reported, as well as how students and instructors can overcome a sense of disorientation due to the pandemic emergency.

5 Discussions and implications

ITSs have been further enhanced with artificial intelligence related technologies (Mousavinasab et al., 2018), which made them an important tool to enable personalized learning and transform teaching methods, curriculum forms and learning environments. This review study aims at systematically analyzing and synthesizing the studies conducted during 2011–2022 and examined the effectiveness of ITS using social experiment method. The findings indicated that the number of studies is slightly increasing since 2011, reflecting an increasing interest among researchers and practitioners towards using social experiments to evaluate ITSs. For the study context, there is a regional "intelligent" divide in the application of ITS, as the distribution of studies was unbalanced geographically and highly focusing on the USA as a context. The most apparent difference of the ITS application in different countries mainly lays in the number of studies conducted. Most of the studies were carried out in the USA, resulting in a diversified ITS application in this country, including using ITS in after school program to enhance learning interaction for math learning (Craig et al., 2013), generating learning tasks for practice for math learning (Jiang et al., 2020),

providing tutoring dialogs throughout learning process (Nye et al., 2018). In contrast, in others countries, there were significantly less forms of ITS application since there were less studies conducted. This result is somehow not consistent with what Nye (2015) indicated that AIED community is increasing and recognizing the importance of designing technologies in the global wide and the digital divide is narrowing. It may because conducting large-scale and long-time-span social experiments is even more complex grounded in social reality, which needs the driving force related to social policies (Forget, 2019). This geographical distribution can also be explained by the statistical data of the national financial investments in AI provided by OECD.AI (2022). The countries where most of the ITS-related social experiment were conducted, also have the most AI-related financial investment. Based on this OECD data and the findings of this present study, it seems that sufficient financial investment is crucial for conducting ITS-related application and social experiment studies. Thus, how to mobilize and share resources, and mitigate this ITS related "intelligent divide" is a challenge for related stakeholders (e.g., policy makers, researchers and practitioners in this field) to address.

The featured functions of the merged ITSs primarily focused on recommendation and tutoring, followed by personalized support, exercise and assessment, personalization, adaptive conversation, and game-based learning. Among various ITSs, Cognitive Tutors were used most extensively and therefore they were the most influential. In addition, complicated ITSs which combine multiple ITSs to process complex tasks were developed and examined. For example, Goel and Joyner (2017) used 100 "Nano tutors" (processed simple tasks) and coordinated them to help students learn AI skills. SKOPE-IT, which combined 2 ITSs, namely, Auto Tutor and ALEKS, was applied to help math-related learning (Nye et al., 2018). This finding responded to and supported by Padadines and Ramírez (2020)'s suggestion that ITS solutions should be more re-usable and take advantages of the existing ITSs as building blocks to save time and costs. These new merging forms of ITSs may indicate the new trend in the principles of designing and developing ITSs in the future.

Regarding the characteristics of the studies in terms of educational context, less studies are conducted in adult learning and kindergartens. It may be because kindergarteners lack the necessary technology or regulatory skills to participate in ITS- supported learning (Casas et al., 2011). For adults, they may have less opportunities of formal, intensive and regular learning compared to K-12 students. For the focused subjects, consistent with Padadines and Ramírez (2020), the current study found that math, language, and science were the primary subjects where ITSs were examined using social experiment method. On the other hand, history, economics, and engineering were less investigated using ITS and social experiment method. However, Ma et al. (2014) found that studies using ITSs in humanities and social science had a significantly higher weighted mean effect size than those which used ITS in math, computer science, physics, literacy, chemistry, and language. Such inconsistency in results suggests the challenging areas of ITS application and potential opportunities for applying and evaluating ITS in humanity and social science subjects, calling

for further investigation in this regard. For experimental design, quasi-experiment and random control trials are the dominating design methods. In contrast, longitudinal study and Randomized Alternative-Treatment Design were rarely used. There are also several studies which did not apply randomization in assigning participants, undermining the conclusion that the observed difference cross groups can be attributed to the treatment (Social experiment, 2008). Consistently with Padadines and Ramírez (2020)'s finding about lacking rigorous evaluation, this current study found that most of the studies did not apply matching techniques to formulate control groups in a solid way, suggesting that researchers should be more aware of rigorous study design in the future so that the conclusions can be more valid and generalizable.

To measure the effectiveness of ITSs, most studies used business-as-usual as benchmarking for comparison. Some of them used ITS as vehicles that carries different types of educational interventions. ITS that carried alternative treatments were used as benchmarking for evaluating the focused instructional treatments. The focus of this method usually is on evaluating the sub-components of ITS, which is beneficial for overcoming the Blackbox effect-related disadvantages of the social experiment method (Peck, 2017) and help improve ITS in a more specific way. The blended form with ITS and human tutors was also used as a benchmark. It responded the conclusion of a meta-analysis conducted by the U.S. Department of Education (2010), where blended form is better than pure human method or pure online method. It is possible that the effectiveness of solutions that combine both human and machine are better than the solutions that only has ITSs.

For the impacts of ITS, most of the reviewed studies reported positive results. Learning performance is the primary measured outcome, consistent with Ma et al. (2014)'s and Mousavinasab et al. (2018)'s findings. The effectiveness of ITSs on students' help seeking, learning interest, engagement, attitude and confidence is relatively less examined, calling for future investigation in this regard. Based on this finding, we suggest to shift the focus of the application and evaluation of ITS from outcome-related cognitive constructs (e.g., learning performance) to the process-related constructs (e.g., engagement, interests, etc.) and non-cognitive constructs (attitudes, confidence, etc.) and from the level of individual learning to a broader social context where learning occurred. The identified mixed types of results regarding the effectiveness of ITS on learning performance and engagement are crucial issues to address. Ma et al. (2014) pointed out that moderator factors (e.g., ITS characteristics), contextual factors (e.g., educational setting) and research design can affect the actual effects of ITS on the outcome variables. However, most of the studies in this current review did not measure these factors or fully control their potential effects in the analysis, which may further lead to mixed or conflicting results. We therefore suggest that future studies examine the influence of ITS from a broader social science perspective, and consider the moderating and mediating factors when designing an experiment and conducting data analysis.

Our findings also identified a set of challenges when applying social experiment to evaluate ITSs, including students' limited task involvement, individual

differences, limited resources, methodological and contextual challenges related to social experiment itself and adaptation of teachers to this ITS-based new form of learning and curriculum. Failing to consider and address such challenges in advance can be the causes of the mixed types of results regarding to the effectiveness of ITSs. Future work can start from addressing these challenges. Moreover, close collaboration involving subject matter experts (social science experts, statistician, etc.) from different disciplines is needed to address these challenges and ensure the success throughout the design, development, application, experimentation and evaluation stages for examining the effectiveness of ITS.

Our findings further imply that researchers in this filed should consider how to increase students' task involvement when applying ITSs, how to handle students' individual differences and contextual features, and use the emerging new methods to design the experiment and analyze the data so that it is possible to accurately measure the effects caused by different components of ITS. Related developers may get inspired by the merging ITSs and related features for future development; researchers, practitioners, and policy makers should be aware of the digital divide across countries and regions, and share experiences and resources to advance the application of ITS in the global-wide.

6 Conclusions and limitations

This systematic review depicted a complicated landscape of the primary studies during 2011–2022 that examined ITS with social experiment method. It contributes to the literature through identifying the latest trends and challenges, and potential factors that can explain the mixed results regarding the effectiveness of ITSs in real and natural educational contexts. Overall, our findings confirmed that ITS can be very powerful to support teaching and learning. However, through the lens of social experiment, it also implies that technology itself cannot guarantee the success of ITS application. The complicated contextual and social factors in real educational fields can influence the observed effectiveness of ITSs. For study methods, this study suggests applying randomization in participants assignment, using matching technique to form comparable control groups, and conducting more rigorous analysis to control the effects of confounding factors. In addition, more attention should be put on the non-cognitive and process-related outcomes.

It should be noted that this study has some limitations that should be acknowledged. For instance, the results of this study are limited by the used search keywords and the selected electronic databases and time span. However, despite these limitations, this study provided solid grounds for investigating the use of social experiment methods to assess the effectiveness ITSs. Future work could focus on the challenges and future research directions reported in this present study to provide more insights about this research topic.

Appendix

Table 14 The ITS articles included for analysis

#	Year	Authors	Title
1	2011	Ward, W.; Cole, R.; Bolanos, D.; Buchenroth-Martin, C.; Svirsky, E.; Van Vuuren, S.; Weston, T.; Zheng, J.; Becker, L	My Science Tutor: A Conversational Multi-Media Virtual Tutor for Elementary School Science
2	2011	Roll, Ido; Aleven, Vincent; McLaren, Bruce M.; Koedinger, Kenneth R	Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system
3	2012	Kegel, Cornelia A. T.; Bus, Adriana G	Online Tutoring as a Pivotal Quality of Web-Based Early Literacy Programs
4	2012	Wijekumar, K. K., Meyer, B. J., & Lei, P	Large-scale randomized controlled trial with 4th graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension
5	2013	Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., & Weston, T	My Science Tutor: A Conversational Multimedia Virtual Tutor
6	2013	Huang, X., Craig, S. D., Xie, J., Graesser, A. C., Okwumabua, T., Cheney, K. R., & Hu, X	The Relationship between Gender, Ethnicity, and Technology on the Impact of Mathematics Achievement in an After-School Program
7	2013	Craig, S. D., Hu, X., Graesser, A. C., Bargagliotti, A. E., Sterbinsky, A., Cheney, K. R., & Okwumabua, T	The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors
8	2013	Fang, N., & Guo, Y	A Web-Based Interactive Intelligent Tutoring System for Undergraduate Engineering Dynamics
9	2013	Lee, H. S., Anderson, J. R., Berman, S. R., Ferris-Glick, J., Joshi, A., Nixon, T., & Ritter, S	Exploring Optimal Conditions of Instructional Guidance in an Algebra Tutor
10	2013	Butcher, K. R., & Aleven, V	Using Student Interactions to Foster Rule-Diagram Mapping During Problem Solving in an Intelligent Tutoring System
11	2013	Wijekumar, K. K., Meyer, B. J., & Lei, P	High-fidelity implementation of web-based intelligent tutoring system improves fourth and fifth graders content area reading comprehension
12	2014	Wijekumar, K., Meyer, B. J., Lei, P. W., Lin, Y. C., Johnson, L. A., Spielvogel, J. A., Shurmatz, K. M., 10 Ray, M., & Cook, M	Multisite Randomized Controlled Trial Examining Intelligent Tutoring of Structure Strategy for Fifth-Grade Readers
13	2014	Feng, M., Roschelle, J., Heffernan, N., Fairman, J., & Murphy, R	Implementation of an Intelligent Tutoring System for Online Homework Support in an Efficacy Trial
14	2014	Pane, J. F., Mcffrey, D. F., Karam, R	Effectiveness of Cognitive Tutor Algebra I

Table 14 (continued)

#	Year	Authors	Title
15	2016	Bartelet, D., Ghysels, J., Groot, W., Haelermans, C., & Maassen van den Brink, H	The Differential Effect of Basic Mathematics Skills Homework via a Web-Based Intelligent Tutoring System Across Achievement Subgroups and Mathematics Domains: A Randomized Field Experiment
16	2016	Chang, Y. H., Chen, Y. Y., Chen, N. S., Lu, Y. T., & Fang, R. J	Yet another adaptive learning management system based on Felder and Silverman's learning styles and Mashup
17	2017	Zhang, B., & Jia, J	Evaluating an Intelligent Tutoring System for Personalized Math Teaching
18	2017	Long, Y., & Aleven, V	Enhancing learning outcomes through self-regulated learning support with an Open Learner Model
19	2017	Wetzel, J., VanLehn, K., Butler, D., Chaudhari, P., Desai, A., Feng, J., … & van de Sande, B	The design and development of the dragoon intelligent tutoring system for model construction: lessons learned
20	2017	Goel, A. K., & Joyner, D. A	Using AI to Teach AI; Lessons from an Online AI Class
21	2017	Klaveren, C., Vonk, S., & Cornelisz, I. (	The effect of adaptive versus static practicing on student learning—evidence from a randomized field experiment
22	2017	Yao, C. B	Constructing a User-Friendly and Smart Ubiquitous Personalized Learning Environment by Using a Context-Aware Mechanism
23	2018	Nye, B. D., Pavlik, P. I., Windsor, A., Olney, A. M., Hajeer, M., & Hu, X	SKOPE-IT (Shareable Knowledge Objects as Portable Intelligent Tutors): overlaying natural language tutoring on an adaptive learning system for mathematics
24	2018	Bernacki, M. L., & Walkington, C	The Role of Situational Interest in Personalized Learning
25	2019	Spichtig, A. N., Gehsmann, K. M., Pascoe, J. P., & Ferrara, J. D	The impact of adaptive, web-based, Scaffolded silent reading Instruction on the reading Achievement of students in Grades 4 and 5
26	2019	Cung, B., Xu, D., Eichhorn, S., & Warschauer, M	Getting Academically Underprepared Students Ready through College Developmental Education: Does the Course Delivery Format Matter?
27	2019	Alshammari, M. T., & Qtaish, A	Effective adaptive e-learning systems according to learning style and knowledge level
28	2020	Watkins, P., C., Caporal, J., Merville, C., Kouider, S., & Dehaene, S	Accelerating reading acquisition and boosting comprehension with a cognitive science-based tablet training

Table 14 (continued)

#	Year	Authors	Title
29	2020	Jiang, Y., Almeda, M., Kai S., Baker, R.S., Ostrow, K., Inventado, P.S., Scupelli P	Single template vs. Multiple templates: Examining the effects of problem format on performance
30	2020	Eryilmaz, M., & Adabashi, A	Development of an Intelligent Tutoring System Using Bayesian Networks and Fuzzy Logic for a Higher Student Academic Performance
31	2020	Treceño-Fernández, D., Calabia-Del-Campo, J., Bote-Lorenzo, M. L., Gómez-Sánchez, E., Luis-García, R., & Alberola-López, C	Integration of an intelligent tutoring system in a magnetic resonance simulator for education: Technical feasibility and user experience
32	2020	Hickey, D. T., Robinson, J., Fiorini, S., & Feng, Y	Internet-based alternatives for equitable preparation, access, and success in gateway courses
33	2020	Baker, D. L., Ma, H., Polanco, P., Conry, J. M., Kamata, A., Al Otaiba, S., Ward, W., & Cole, R	Development and promise of a vocabulary intelligent tutoring system for Second-Grade Latinx English learners
34	2020	Ahmed, H., Wilson, A., Mead, N., Noble, H., Richardson, U., Wolpert, M. A., & Goswami, U	An Evaluation of the Efficacy of GraphoGame Rime for Promoting English Phonics Knowledge in Poor Readers
35	2020	Wijekumar, K., Meyer, B. J., Lei, P., Beerwinkle, A. L., & Joshi, M	Supplementing teacher knowledge using web-based Intelligent Tutoring System for the Text Structure Strategy to improve content area reading comprehension with fourth- and fifth-grade struggling readers
36	2020	Tacoma, S., Drijvers, P., & Jeuring, J	Combined inner and outer loop feedback in an intelligent tutoring system for statistics in higher education
37	2021	Mousavi, A., Schmidt, M., Squires, V., & Wilson, K	Assessing the Effectiveness of Student Advice Recommender Agent (SARA): the Case of Automated Personalized Feedback
38	2021	Troussas, C., Krouska, A., & Sgouropoulou, C	A Novel Teaching Strategy Through Adaptive Learning Activities for Computer Programming
39	2022	Capone, R., de Falco, M., Lepore, M	The Impact of Covid-19 Pandemic on Undergraduate Students: the Role of an Adaptive Situation-Aware Learning System
40	2022	del Olmo-Muñoz1, J., González-Calero, J., Diago, P. D., Arnau, D., Arevalillo-Herráez, M	Using intra-task flexibility on an intelligent tutoring system to promote arithmetic problem-solving proficiency

Acknowledgements The current study isfunded by Guangdong University Online Open Course Committee (广东省本科高校在线开放课程指导委员会, Grant Number: 2022ZXKC397) and China Institue of Education and Social Development (中国教育与社会发展研究院, Grant Number: Gb2021013).

Data availability The datasets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Ahmed, H., Wilson, A., Mead, N., Noble, H., Richardson, U., Wolpert, M. A., & Goswami, U. (2020). An evaluation of the efficacy of GraphoGame rime for promoting English phonics knowledge in poor readers. *Frontiers in Education, 5*, 132. https://doi.org/10.3389/feduc.2020.00132

Alabdulhadi, A., & Faisal, M. (2021). Systematic literature review of STEM self-study related ITSs. *Education and Information Technologies, 26*(2), 1549–1588. https://doi.org/10.1007/s10639-020-10315-z

Alshammari, M. T., & Qtaish, A. (2019). Effective adaptive e-learning systems according to learning style and knowledge level. *Journal of Information Technology Education*, *18*, 529–547. Retrieved December 29th, 2022 from http://www.jite.org/documents/Vol18/JITEv18ResearchP529-547Alshammari5698.pdf

Atun, H. (2020). Intelligent tutoring systems (its) to improve reading comprehension: a systematic review. *Journal of Teacher Education and Lifelong Learning, 2*(2), 77–89. Retrieved December 29th, 2022 from https://dergipark.org.tr/en/pub/tell/issue/58491/757329

Baker, D. L., Ma, H., Polanco, P., Conry, J. M., Kamata, A., Al Otaiba, S., Ward, W., & Cole, R. (2020). Development and promise of a vocabulary intelligent tutoring system for second-grade Latinx English learners. *Journal of Research on Technology in Education, 53*(2), 223–247. https://doi.org/10.1080/15391523.2020.1762519

Bartelet, D., Ghysels, J., Groot, W., Haelermans, C., & Maassen van den Brink, H. (2016). The differential effect of basic mathematics skills homework via a web-based intelligent tutoring system across achievement subgroups and mathematics domains: A randomized field experiment. *Journal of Educational Psychology, 108*(1), 1–20. https://doi.org/10.1037/edu0000051

Bernacki, M. L., & Walkington, C. (2018). The role of situational interest in personalized learning. *Journal of Educational Psychology, 110*(6), 864–881. https://doi.org/10.1037/edu0000250

Burns, H.L., & Capps, C.G. (1988). Foundations of intelligent tutoring systems: An introduction. In M. C. Polson & J. J. Richardson (Eds.), *Foundations of intelligent tutoring systems* (pp. 1–19). Lawrence Erlbaum. https://www.taylorfrancis.com/chapters/mono/10.4324/9780203761557-6/foundations-intelligent-tutoring-systems-introduction-martha-polson-jeffrey-richardson

Butcher, K. R., & Aleven, V. (2013). Using student interactions to foster rule–diagram mapping during problem solving in an intelligent tutoring system. *Journal of Educational Psychology, 105*(4), 988–1009. https://doi.org/10.1037/a0031756

Capone, R., De Falco, M., & Lepore, M. (2022). The Impact of Covid-19 Pandemic on Undergraduate Students: the Role of an Adaptive Situation-Aware Learning System. In 2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA) (pp. 154-161). IEEE.

Carter, E. E. (2014). *An intelligent debugging tutor for novice computer science students*. [Doctoral dissertation, Lehigh University]. ProQuest Dissertations & Theses Global. Retrieved December 29th, 2022 from https://www.proquest.com/dissertations-theses/intelligent-debugging-tutor-novice-computer/docview/1540757322/se-2

Casas, I., Goodman, P.S., Pelaez, E. (2011). *On the design and use of a cognitive tutoring system in the math classroom* [Paper presentation]. 2011 IEEE International Conference on Technology for Education, Chennai, Tamil Nadu India. Retrieved December 29th, 2022 from https://ieeexplore.ieee.org/document/6004354

Chang, Y. H., Chen, Y. Y., Chen, N. S., Lu, Y. T., & Fang, R. J. (2016). Yet another adaptive learning management system based on Felder and Silverman's learning styles and Mashup. *Eurasia Journal of Mathematics, Science and Technology Education*, *12*(5), 1273–1285. http://www.ejmste.com/ms.aspx?kimlik=10.12973/eurasia.2016.1512a

Churi, P. P., Joshi, S., Elhoseny, M., & Omrane, A. (Eds.). (2022). *Artificial intelligence in higher education: A practical approach* (1st ed.). CRC Press. https://doi.org/10.1201/9781003184157

Colby, B. R. (2017). *A comparative literature review of intelligent tutoring systems from 1990–2015*. [Master's thesis, Brigham Young University]. Scholars Archive. Retrieved December 29th, 2022 from https://scholarsarchive.byu.edu/etd/7239/

Corbett, A. T., Koedinger, K., & Hadley, W. S. (2001). Cognitive tutors: From the research classroom to all classrooms. In P. S. Goodman (Ed.), *Technology enhanced learning: Opportunities for change* (pp. 235–263). Lawrence Erlbaum Associates Publishers.

Craig, S. D., Hu, X., Graesser, A. C., Bargagliotti, A. E., Sterbinsky, A., Cheney, K. R., & Okwumabua, T. (2013). The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors. *Computers & Education, 68*, 495–504. https://doi.org/10.1016/j.compedu.2013.06.010

Crow, T., Luxton-Reilly, A., & Wuensche, B. (2018, January 30- February 2). *Intelligent tutoring systems for programming education: a systematic review* [Paper presentation]. 20th Australasian Computing Education Conference, Brisbane, Queensland, Australia.

Cuéllar-Rojas, O. A., Hincapié, M., Contero, M., & Güemes-Castorena, D. (2021). Intelligent tutoring system: A bibliometric analysis and systematic literature review. *Research Square*. Advance online publication. https://doi.org/10.21203/rs.3.rs-673038/v1

Cung, B., Xu, D., Eichhorn, S., & Warschauer, M. (2019). Getting academically underprepared students ready through college developmental education: Does the course delivery format matter? *American Journal of Distance Education, 33*(3), 178–194. https://doi.org/10.1080/08923647.2019.1582404

del Olmo-Muñoz, J., González-Calero, J. A., Diago, P. D., Arnau, D., & Arevalillo-Herráez, M. (2022). Intelligent tutoring systems for word problem solving in COVID-19 days: could they have been (part of) the solution? ZDM–Mathematics Education, 1–14.

Eryilmaz, M., & Adabashi, A. (2020). Development of an intelligent tutoring system using bayesian networks and fuzzy logic for a higher student academic performance. *Applied Sciences, 10*(19), 6638. https://doi.org/10.3390/app10196638

Fang, N., & Guo, Y. (2013). *A web-based interactive intelligent tutoring system for undergraduate engineering dynamics* [Paper presentation]. 2013 IEEE Frontiers in Education Conference, Oklahoma City, USA.

Feng, M., Roschelle, J., Heffernan, N., Fairman, J., & Murphy, R. (2014). *Implementation of an intelligent tutoring system for online homework support in an efficacy trial* [Paper presentation]. 12th International Conference on Intelligent Tutoring Systems, Verlag, Berlin, Heidelberg.

Feng, S., Magana, A. J., & Kao, D. (2021). *A systematic review of literature on the effectiveness of intelligent tutoring systems in STEM* [Paper presentation]. 2021 IEEE Frontiers in Education Conference (FIE), Lincoln, NE, USA.

Fienberg S. E., Singer B., Tanur J.M. (1985). Large-Scale Social Experimentation in the United States. In A. C. Atkinson & S. E. Fienberg (Eds), *A Celebration of Statistics* (pp. 287–326). Springer, New York. https://doi.org/10.1007/978-1-4613-8560-8_12

Forget. (2019). *Experiments in Society: Framing social experiments at the boundary between social work and sociology* [Paper presentation]. Centre for the History of Political Economy (CHOPE), Durham, North Carolina, USA.

Goel, A. K., & Joyner, D. A. (2017). Using AI to teach AI: Lessons from an online AI class. *AI Magazine, 38*(2), 48–59. https://doi.org/10.1609/aimag.v38i2.2732

Greenberg, D., & Shroder, M. (2004). *The digest of social experiments*. Urban Institute Press.

Guha, M. (2008), International encyclopedia of the social sciences (2nd edition). *Reference Reviews, 22*(7), 17–19. https://doi.org/10.1108/09504120810905060

Harris, J. D., Quatman, C. E., Manring, M. M., Siston, R. A., & Flanigan, D. C. (2014). How to write a systematic review. *The American Journal of Sports Medicine, 42*(11), 2761–2768. https://doi.org/10.1177/0363546513497567

Hickey, D. T., Robinson, J., Fiorini, S., & Feng, Y. (2020). Internet-based alternatives for equitable preparation, access, and success in gateway courses. *The Internet and Higher Education, 44*, 100693. https://doi.org/10.1016/j.iheduc.2019.100693

Huang, X., Craig, S. D., Xie, J., Graesser, A. C., Okwumabua, T., Cheney, K. R., & Hu, X. (2013). *The relationship between gender, ethnicity, and technology on the impact of mathematics achievement in an after-school program* [Paper presentation]. Society for Research on Educational Effectiveness Spring 2013, Washington, D.C., USA.

Jiang, Y., Almeda, M., Kai, S., Baker, R. S., Ostrow, K., Inventado, P. S., & Scupelli, P. (2020). *Single template vs. multiple templates: Examining the effects of problem format on performance* [Paper presentation]. The 14th *International Conference on the Learning Sciences*, Nashville, Tennessee.

Kegel, C. A., & Bus, A. G. (2012). Online tutoring as a pivotal quality of web-based early literacy programs. *Journal of Educational Psychology, 104*(1), 182–192. https://doi.org/10.1037/A0025849

Keleş, A., Ocak, R., Keleş, A., & Gülcü, A. A. (2009). ZOSMAT: Web-based intelligent tutoring system for teaching–learning process. *Expert Systems with Applications, 36*(2), 1229–1239. https://doi.org/10.1016/j.eswa.2007.11.064

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering (EBSE 2007–001).* Retrieved December 29th, 2022 from https://www.researchgate.net/publication/302924724_Guidelines_for_performing_Systematic_Literature_Reviews_in_Software_Engineering

Klaveren, C., Vonk, S., & Cornelisz, I. (2017). The effect of adaptive versus static practicing on student learning-evidence from a randomized field experiment. *Economics of Education Review, 58*, 175–187. https://doi.org/10.1016/j.econedurev.2017.04.003

Koedinger, K. R., & Aleven, V. (2016). An interview reflection on "Intelligent tutoring goes to school in the big city". *International Journal of Artificial Intelligence in Education, 26*(1), 13–24. https://link.springer.com/article/10.1007/s40593-015-0082-8

Lee, H. S., Anderson, J. R., Berman, S. R., Ferris-Glick, J., Joshi, A., Nixon, T., & Ritter, S. (2013). *Exploring Optimal Conditions of Instructional Guidance in an Algebra Tutor* [Paper presentation]. Society for Research on Educational Effectiveness Fall 2013, Washington, D.C., USA.

Li, K. C., & Wong, B. T. M. (2021). Features and trends of personalised learning: A review of journal publications from 2001 to 2018. *Interactive Learning Environments, 29*(2), 182–195. https://doi.org/10.1080/10494820.2020.1811735

Long, Y., & Aleven, V. (2017). Enhancing learning outcomes through self-regulated learning support with an open learner model. *User Modeling and User-Adapted Interaction, 27*(1), 55–88. https://doi.org/10.1007/s11257-016-9186-6

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education.

Ma, W., Adesope, O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918. https://doi.org/10.1037/a0037123

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement: elaboration and explanation. *BMJ: British Medical Journal*, *349*, g7647. https://doi.org/10.1136/bmj.g7647

Mousavi, A., Schmidt, M., Squires, V., & Wilson, K. (2021). Assessing the effectiveness of student advice recommender agent (SARA): The case of automated personalized feedback. *International Journal of Artificial Intelligence in Education, 31*(3), 603–621. https://doi.org/10.1007/s40593-020-00210-6

Mousavinasab, E., Zarifsanaiey, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2018). Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, *29*(1), 142–163.https://doi.org/10.1080/10494820.2018.1558257

Neagu, L. M., Rigaud, E., Travadel, S., Dascalu, M., & Rughinis, R. V. (2020, June 8–12). *Intelligent tutoring systems for psychomotor training–a systematic literature review* [Paper presentation]. International Conference on Intelligent Tutoring Systems 2020, Athens, Greece.

Nwana, H. S. (1990). Intelligent tutoring systems: An overview. *Artificial Intelligence Review, 4*, 251–277. https://doi.org/10.1007/BF00168958

Nye, B. D. (2015). Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. *International Journal of Artificial Intelligence in Education, 25*, 177–203. https://doi.org/10.1007/s40593-014-0028-6

Nye, B. D., Pavlik, P. I., Windsor, A., Olney, A. M., Hajeer, M., & Hu, X. (2018). SKOPE-IT (Shareable Knowledge Objects as Portable Intelligent Tutors): Overlaying natural language tutoring on an adaptive learning system for mathematics. *International Journal of STEM Education, 5*(1), 1–20. https://doi.org/10.1186/s40594-018-0109-4

OECD.AI (2022), Visualisations powered by JSI using data from Preqin, Retrieved December 29th, 2022 from www.oecd.ai

Paladines, J., & Ramírez, J. (2020). A systematic literature review of intelligent tutoring systems with dialogue in natural language. *IEEE Access, 8*, 164246–164267. https://doi.org/10.1109/ACCESS.2020.3021383

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, 36(2), 127–144.

Peck, L. R. (Ed.). (2017). *Social experiments in practice: The what, why, when, where, and how of experimental design and analysis: New Directions for Evaluation, Number 152*. John Wiley & Sons.

Riecken, H. W., & Boruch, R. F. (1974). *Social Experimentation: A Method for Planning and Evaluating Social Intervention*. Academic Press.

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

Rolston, H. (2016). On the "why" of social experiments: Some lessons on overcoming barriers from 45 Years of social experiments. *New Directions for Evaluation, 2016*(152), 19–31. https://doi.org/10.1002/ev.20214

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational researcher*, *15*(9), 5–11. Retrieved December 29th, 2022 from https://www.jstor.org/stable/1174711

Social Experiment. (2008). In W. A. Darity, Jr. (Ed.), *International Encyclopedia of the Social Sciences* (2nd ed., Vol. 7, pp. 590–592). Macmillan Reference USA. Retrieved December 29th, 2022 from https://link.gale.com/apps/doc/CX3045302492/WHIC?u=cnbnu&sid=bookmark-WHIC&xid=8e9ed662

Soofi, A. A., & Ahmed, M. U. (2019). A systematic review of domains, techniques, delivery modes and validation methods for intelligent tutoring systems. *International Journal of Advanced Computer Science and Applications, 10*(3), 99–107. https://doi.org/10.14569/IJACSA.2019.0100312

Spichtig, A. N., Gehsmann, K. M., Pascoe, J. P., & Ferrara, J. D. (2019). The impact of adaptive, web-based, scaffolded silent reading instruction on the reading achievement of students in grades 4 and 5. *The Elementary School Journal, 119*(3), 443–467. https://doi.org/10.1086/701705

Tacoma, S., Drijvers, P., & Jeuring, J. (2020). Combined inner and outer loop feedback in an intelligent tutoring system for statistics in higher education. *Journal of Computer Assisted Learning, 37*(2), 319–332. https://doi.org/10.1111/jcal.12491

Treceño-Fernández, D., Calabia-Del-Campo, J., Bote-Lorenzo, M. L., Gómez-Sánchez, E., Luis-García, R., & Alberola-López, C. (2020). Integration of an intelligent tutoring system in a magnetic resonance simulator for education: Technical feasibility and user experience. *Computer Methods and Programs in Biomedicine, 195*, 105634. https://doi.org/10.1016/j.cmpb.2020.105634

Troussas, C., Krouska, A., & Sgouropoulou, C. (2021). A novel teaching strategy through adaptive learning activities for computer programming. *IEEE Transactions on Education, 64*(2), 103–109. https://doi.org/10.1109/TE.2020.3012744

U.S. Department of Education. (2010). *Evaluation evidence-based practices in online learning meta-analysis and review of online learning studies*. Office of Planning, Evaluation, and Policy Development.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., & Weston, T. (2013). My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology, 105*(4), 1115. https://doi.org/10.1037/a0031589

Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S. V., ..., & Becker, L. (2011). My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Transactions on Speech and Language Processing (TSLP)*, *7*(4), 1–29.https://doi.org/10.1145/1998384.1998392

Watkins, P. C., Caporal, J., Merville, C., Kouider, S., & Dehaene, S. (2020). Accelerating reading acquisition and boosting comprehension with a cognitive science-based tablet training. *Journal of Computers in Education, 7*(2), 183–212. https://doi.org/10.1007/s40692-019-00152-6

Wetzel, J., VanLehn, K., Butler, D., Chaudhari, P., Desai, A., Feng, J., ..., & van de Sande, B. (2017). The design and development of the dragoon intelligent tutoring system for model construction: lessons learned. *Interactive Learning Environments*, *25*(3), 361–381https://doi.org/10.1080/10494820.2015.1131167

Wijekumar, K. K., Meyer, B. J., & Lei, P. (2012). Large-scale randomized controlled trial with 4th graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension. *Educational Technology Research and Development*, *60*(6), 987–1013. https://www.jstor.org/stable/23356890

Wijekumar, K. K., Meyer, B. J., & Lei, P. (2013). High-fidelity implementation of web-based intelligent tutoring system improves fourth and fifth graders content area reading comprehension. *Computers & Education, 68*, 366–379. https://doi.org/10.1016/j.compedu.2013.05.021

Wijekumar, K., Meyer, B. J., Lei, P. W., Lin, Y. C., Johnson, L. A., Spielvogel, J. A., Shurmatz, K. M., Ray, M., & Cook, M. (2014). Multisite randomized controlled trial examining intelligent tutoring of structure strategy for fifth-grade readers. *Journal of Research on Educational Effectiveness, 7*(4), 331–357. https://doi.org/10.1080/19345747.2013.853333

Wijekumar, K., Meyer, B. J., Lei, P., Beerwinkle, A. L., & Joshi, M. (2020). Supplementing teacher knowledge using web-based intelligent tutoring system for the text structure strategy to improve content area reading comprehension with fourth-and fifth-grade struggling readers. *Dyslexia, 26*(2), 120–136. https://doi.org/10.1002/dys.1634

Yao, C. B. (2017). Constructing a user-friendly and smart ubiquitous personalized learning environment by using a context-aware mechanism. *IEEE Transactions on Learning Technologies, 10*(1), 104–114. https://doi.org/10.1109/TLT.2015.2487977

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1–27. https://doi.org/10.1186/s41239-019-0171-0

Zhang, B., & Jia, J. (2017). *Evaluating an intelligent tutoring system for personalized math teaching* [Paper presentation]. 2017 international symposium on educational technology, Hong Kong, China.

Authors and Affiliations

Huanhuan Wang[1] · Ahmed Tlili[2] · Ronghuai Huang[1] · Zhenyu Cai[3] · Min Li[1] · Zui Cheng[4] · Dong Yang[2] · Mengti Li[5] · Xixian Zhu[1] · Cheng Fei[2]

Huanhuan Wang
holly.08@live.cn

Ahmed Tlili
ahmed.tlili23@yahoo.com

Ronghuai Huang
huangrh@bnu.edu.cn

Zhenyu Cai
zhenyu_cai@163.com

Min Li
bsxiaomin@163.com

Dong Yang
dydyor@outlook.com

Mengti Li
limengti951001@163.com

Xixian Zhu
eunicezhu997@163.com

Cheng Fei
feicheng@bnu.edu.cn

[1] Faculty of Education, National Engineering Research Center of Cyberlearning and Intelligent Technology (China), Beijing Normal University, Beijing, China

[2] Smart Learning Institute of Beijing Normal University, Beijing, China

[3] Computer-Human Interaction in Learning and Instruction (CHILI) Lab, Écolecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[4] Faculty of Education, Shenzhen University, Shenzhen, China

[5] Educational Informatization Strategy Research Base, Ministry of Education, Beijing, China