Received: 25 January 2023 | Accepted: 17 April 2023

DOI: 10.1111/bjet.13334

REVIEW

British Journal of Educational Technology BERA

Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis

Rong Wu | **Zhonggen Yu**

Faculty of Foreign Studies, Beijing Language and Culture University, Beijing, China

Correspondence
Zhonggen Yu, Faculty of Foreign Studies, Beijing Language and Culture University, 15 Xueyuan Road, Haidian District, Beijing 100083, China.
Email: 401373742@qq.com

Funding information
MOOC of Beijing Language and Culture University, Grant/Award Number: MOOC201902; "Introduction to Linguistics"; "Introduction to Linguistics" of online and offline mixed courses in Beijing Language and Culture University in 2020; Special fund of Beijing Co-construction Project-Research and reform of the "Undergraduate Teaching Reform and Innovation Project" of Beijing higher education in 2020-innovative "multilingual+"excellent talent training system, Grant/Award Number: 202010032003; the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University, Grant/Award Number: 22YCX038; Beijing Language and Culture University Excellent Doctoral Dissertation Cultivation Program Funding Project

Abstract

Artificial intelligence (AI) chatbots are gaining increasing popularity in education. Due to their increasing popularity, many empirical studies have been devoted to exploring the effects of AI chatbots on students' learning outcomes. The proliferation of experimental studies has highlighted the need to summarize and synthesize the inconsistent findings about the effects of AI chatbots on students' learning outcomes. However, few reviews focused on the meta-analysis of the effects of AI chatbots on students' learning outcomes. The present study performed a meta-analysis of 24 randomized studies utilizing Stata software (version 14). The main goal of the current study was to meta-analytically examine the effects of AI chatbots on students' learning outcomes and the moderating effects of educational levels and intervention duration. The results indicated that AI chatbots had a large effect on students' learning outcomes. Moreover, AI chatbots had a greater effect on students in higher education, compared to those in primary education and secondary education. In addition, short interventions were found to have a stronger effect on students' learning outcomes than long interventions. It could be explained by the argument that the novelty effects of AI chatbots could improve learning outcomes in short interventions, but it has worn off in the long interventions. Future designers and educators should make attempt to increase students' learning outcomes by equipping AI chatbots with human-like avatars, gamification elements and emotional intelligence.

KEYWORDS
artificial intelligence chatbots, educational levels, intervention duration, learning outcomes, meta-analysis

Practitioner notes

What is already known about this topic

- In recent years, artificial intelligence (AI) chatbots have been gaining increasing popularity in education.
- Studies undertaken so far have provided conflicting evidence concerning the effects of AI chatbots on students' learning outcomes.
- There has remained a paucity of meta-analyses synthesizing the contradictory findings about the effects of AI chatbots on students' learning outcomes.

What this paper adds

- This study, through meta-analysis, synthesized these recent findings about the effects of AI chatbots on students' learning outcomes.
- This study found that AI chatbots could have a large effect on students' learning outcomes.
- This study found that the effects of AI chatbots were moderated by educational levels and intervention duration.

Implications for practice and/or policy

- AI chatbot designers could make AI chatbots better by equipping AI chatbots with human-like avatars, gamification elements and emotional intelligence
- Practitioners and/or teachers should draw attention to the positive and negative effects of AI chatbots on students.
- Considering the importance of ChatGPT, more research is required to develop a better understanding of the effects of ChatGPT in education.
- More research is needed to examine the mechanisms underlying the effects of AI chatbots on students' learning outcomes.

INTRODUCTION

An introduction to AI chatbots

The rapid development of artificial intelligence (AI) techniques has led to the widespread popularity of AI chatbots (Nguyen et al., 2022). AI chatbots, also known as conversational agents, are a sort of dialogue computer program that could conduct human-like conversations using natural language processing, machine learning, neural networks, information retrieval, deep learning, and other AI techniques (Zhang et al., 2020). AI chatbots with characteristics of interactivity, flexibility and personalization could deal with user requests and then give real-time responses to users via texts, voice or a combination of both (Wu, Lin et al., 2020). There are a variety of AI chatbots available online, such as ChatGPT. Apple's SIRI, Google Assistant and Amazon Lex.

There have been noticeable differences between traditional chatbots and AI chatbots. Traditional chatbots, also called ruled-based chatbots, could merely accomplish a specific task (Shumanov & Johnson, 2021) and select responses from predefined knowledge bases (Luo et al., 2022). However, traditional chatbots still suffer from many limitations, such as the incapacity to understand user intentions and provide personalized guidance (Hwang et al., 2022). AI chatbots could not only save user inputs but also, more importantly, learn from previous user inputs and enable improved interaction (Nguyen et al., 2022). In addition,

AI chatbots could imitate human cognitive functions, including problem-solving, contextualized reasoning, predicting, planning and decision-making (Nadarzynski et al., 2019).

AI chatbots could be generally classified into three types: machine learning-based chatbots, natural language processing-based chatbots and hybrid chatbots. Traditional machine learning-based chatbots could grasp user intent, filter useless information and provide useful guidance by using machine learning algorithms. Machine learning-based chatbots often need a large amount of data for training. However, it might be challenging to find suitable databases for chatbot training (Adamopoulou & Moussiades, 2020). Natural language processing-based chatbots using advanced machine learning algorithms could not only learn from users' previous inputs but also recognize, infer and use human languages. In addition, they could conduct more tasks such as intent recognition, sentiment analysis, and emotion detection (Zhang et al., 2022). Hybrid chatbots have been launched by combining both AI algorithms and rule-based logic. Similar to ruled-based chatbots, hybrid chatbots could solve specific problems. They could also mimic human cognitive functions and deal with more complex problems (Haristiani, 2019). Therefore, AI chatbots have opened up the possibility of success, especially in educational development (Jeon, 2022).

The use of AI chatbots in education

AI chatbots have emerged as new educational tools. Traditional educational tools merely offer access to knowledge, while AI chatbots could take on the roles of partners, assistants and even mentors in learning environments (Fidan & Gencel, 2022; Wollny et al., 2021). AI chatbots could help learners practice languages through interactions like human partners (Lee et al., 2022; Liu, Liao et al., 2022). In addition, AI chatbots could provide learners with personalized learning content, giving feedback and guidance on a one-to-one basis (Fidan & Gencel, 2022). AI chatbots could act as an assistant to help teachers make dynamic assessments of each student (Jeon, 2021), reducing teachers' workload, burden and pressure. Due to their potential, AI chatbots have been widely used in various academic disciplines such as mathematics (Yin et al., 2021), psychology (Lin & Chang, 2020), medicine (Lee et al., 2022), and language (Kim, 2019).

A growing body of research has explored the possible effects of AI chatbots in education (Jeon, 2022). Evidence showed that AI chatbots could promote students' academic performance (Kim, 2018a; Vázquez-Cano et al., 2021), inspire learning interest (Wambsganss et al., 2021), and boost learning motivation (Chien et al., 2022; Kim, 2018b), engagement (Ruan et al., 2021), and learning self-efficacy (Yin et al., 2021). Besides, it was found that students could relieve learning anxiety and stress while using AI chatbots in the learning process (Klos et al., 2021; Terblanche et al., 2022). However, recent studies found that there was no significant difference between the chatbot group and the control group with respect to learning engagement (Liu, Liao et al., 2022), confidence (Han et al., 2022), motivation (Kumar, 2021), and performance (Yin et al., 2021). Therefore, it is important to note that the effects of AI chatbots may be more complex than once thought.

Previous reviews

As a great deal of empirical research provided conflicting evidence concerning the effects of AI chatbots, an important issue that has emerged might be to provide a comprehensive summary and synthesis of conflicting results. Several systematic reviews of the roles of chatbots in education have been undertaken. Wollny et al. (2021), for example, reviewed 74 publications and heightened the beneficial effects of chatbots on students' skills and motivation. In

a systematic review, chatbots were found to have a significant impact on student learning achievement and subjective satisfaction (Kuhail et al., 2022). However, chatbots might lead learners to experience the novelty effect and negative emotions, which would be detrimental to their learning outcomes (Pérez et al., 2020). Another literature review augured that the use of chatbots could facilitate students' language learning, but possibly increase students' extraneous cognitive load (Huang et al., 2022).

The existing research provided a narrative synthesis of evidence assessing the effects of chatbots in education. Nevertheless, there has been still a need to conduct a more comprehensive review. Previous reviews undertook qualitative content analyses of the effects of chatbots in education. They presented and described the effects of chatbots in a systematic and detailed way. Nevertheless, they did not examine the extent of the effects of AI chatbots and provide quantitative evidence for their claims and arguments. They may not avoid the subjectivity inherent in the narrative reviews (Garrido et al., 2018).

Meta-analyses may offer an effective way to avoid the subjectivity inherent in narrative reviews (Pickering & Byrne, 2014). Meta-analyses could integrate primary data gathered from numerous empirical research, allowing researchers to quantify the extent of the effect of an intervention (Bai et al., 2020). Using this approach, researchers could arrive at a more robust and convincing conclusion about the extent of the effect of an intervention (Garrido et al., 2018). Nevertheless, there is a surprising paucity of meta-analyses synthesizing the contradictory findings about the effects of AI chatbots on students' learning outcomes. This study sought to synthesize empirical findings about the effects of AI chatbots on students' learning outcomes in terms of performance, motivation, self-efficacy, interest, anxiety, and perceived value of learning. Further, this study undertook moderator analyses to shed light on how educational levels and intervention duration may impact the effects of AI chatbots on students' learning outcomes.

Drawing on an integrated analytic framework for investigating learning outcomes (Salas-Pilco, 2020), this meta-analysis tried to answer the following research questions: (1) Do AI chatbots greatly improve learning outcomes in terms of performance, motivation, self-efficacy, interest, anxiety, and perceived value of learning? (2) Do educational levels influence learning outcomes? (3) Does the duration of using AI chatbots influence learning outcomes?

LITERATURE REVIEW

Learning performance

Learning performance was operationally defined as the extent to which students could gain and apply valuable knowledge and skills in AI chatbot-based learning. Learning knowledge and skills are generally associated with intellectual outcomes (Salas-Pilco, 2020). Much of the current literature on AI chatbots found improved learning performance in terms of various aspects, including vocabulary acquisition (Kim, 2018b; Ruan et al., 2021), listening tests (Chien et al., 2022; Kim, 2018a), speaking tests (Lin & Mubarok, 2021), writing skills (Lin & Chang, 2020), grammar skills (Nghi et al., 2019), public health knowledge (Lee et al., 2022), cultural knowledge (Mageira et al., 2022), nursing concepts (Chang et al., 2022) and instructional technology knowledge (Fidan & Gencel, 2022). However, recent research revealed that there was no significant difference in learning performance between the chatbot group and the control group (Yin et al., 2021). Given the conflicting findings, we proposed the following hypothesis:

> **H1:** The use of AI chatbots could significantly improve learning performance at the 0.05 level.

Learning motivation

Learning motivation was operationally defined as learners' desire to participate in various learning activities for external rewards or the inherent feelings of satisfaction and enjoyment. Learning motivation is a key component of affective-emotional outcomes (Salas-Pilco, 2020). Several studies demonstrated the positive influence of AI chatbots on learning motivation. Evidence suggested that AI chatbots capable of providing goal-oriented feedback could motivate students to practice English speaking skills (Han, 2020) and listening skills (Chien et al., 2022). Moreover, students using AI chatbots had a higher level of motivation to learn English vocabulary (Kim, 2018b) and knowledge of public health (Lee et al., 2022). Therefore, we proposed the following hypothesis:

H2: The use of AI chatbots could significantly increase learning motivation at the 0.05 level.

Learning self-efficacy

Learning self-efficacy, in this study, was defined as learners' confidence in their capacities to perform behaviours aiming at achieving academic success. Learning self-efficacy is a specific construct of self-efficacy (Kuo et al., 2021). Learning self-efficacy is a significant contributory factor to social–emotional outcomes (Salas-Pilco, 2020). Various findings have been delivered regarding the influence of AI chatbots on learning self-efficacy. Korean college students who used AI chatbots to learn English vocabulary had significantly higher learning self-efficacy than those who did not (Kim, 2018b). Similarly, Lee et al. (2022) found that it might be more effective in using AI chatbots to enhance students' learning self-efficacy in the review process. The use of the chatbot system could also lead to a significant increase in learning self-efficacy among nursing students (Chang et al., 2022). However, Chinese university students did not show a higher level of learning self-efficacy in the chatbot group than in the control group (Yin et al., 2021). Considering the inconsistent findings, we proposed the following hypothesis:

H3: The use of AI chatbots could significantly enhance learning self-efficacy at the 0.05 level.

Learning interest

Learning interest was understood as a learner's disposition to engage in particular learning activities for a certain amount of time (Tsai et al., 2018). Learning interest is one of the key aspects of social–emotional outcomes (Salas-Pilco, 2020). Some quantitative intervention research reported that AI chatbots could inspire and sustain students' learning interests. College students and secondary school students using AI chatbots showed more interest in English learning, compared to those without AI chatbots (Han, 2020; Kim, 2018b). Feedback from AI chatbots could be more effective to increase pre-service teachers' learning interest than online human-written feedback (Fidan & Gencel, 2022). Moreover, AI chatbots could be conducive to sustaining primary school students' situational interest in extensive reading (Liu, Liao et al., 2022). The chatbot-based system could offer affective feedback anytime and anywhere, helping greatly to sustain students' interest (Yin et al., 2021). Therefore, we proposed the following hypothesis:

H4: The use of AI chatbots could significantly increase learning interest at the 0.05 level.

Anxiety

Anxiety, in this study, is operationally defined as learners' negative emotions of tension, unease and self-doubt due to learning difficulty and pressure. Emotions, including anxiety, are strongly tied to social–emotional outcomes (Salas-Pilco, 2020). The use of AI chatbots was found to relieve learners' anxiety. AI chatbots could reduce the anxiety levels of secondary school students in foreign language learning (Han, 2020). University students using the chatbot-delivered intervention reported significantly less anxiety and depression than the control group (Liu, Peng et al., 2022). In contrast to earlier findings, Klos et al. (2021) demonstrated that there was a significant reduction in anxiety symptoms among the chatbot group. Nevertheless, there was no significant difference in anxiety symptoms between the chatbot group and the control group. Overall, the use of AI chatbots may relieve students' anxiety in the learning environment. Therefore, we proposed the following hypothesis:

H5: The use of AI chatbots could significantly relieve anxiety at the 0.05 level.

Perceived value of learning

Learners' perceived value of learning was operationally defined as learners' perception of the value of learning contents and activities in technology-based learning in this study. Social–emotional outcomes are intimately related to the perceived value of learning (Salas-Pilco, 2020). The existing evidence indicated that AI chatbots could plausibly increase the perceived value of learning. For example, university students perceived the value of absorbing and acquiring knowledge in the chatbot-based learning environment (Yin et al., 2021). Fidan and Gencel (2022) also found that pre-service teachers perceived the value of learning knowledge after using the chatbot-based feedback system. Similarly, AI chatbots could help primary school students understand the themes of stories, increasing their perceived value of learning (Liu, Liao et al., 2022). Therefore, we proposed the following hypothesis:

H6: The use of AI chatbots could significantly increase the perceived value of learning at the 0.05 level.

Educational levels

The existing body of research suggested that AI chatbots could affect learning outcomes at different educational levels. AI chatbots could exert a powerful effect on learning outcomes at the primary and secondary school levels. For example, the primary school students using a smart chatbot performed significantly better on tests than those in the control group (Hwang et al., 2022). Students in a Greek secondary school also reported that AI chatbots could help them to learn cultural content and language (Mageira et al., 2022). Similarly, AI chatbots could have a significant effect on students in higher education. Educational chatbots could improve university students' learning performance in team-based projects (Kumar, 2021). Therefore, we proposed the following hypothesis:

H7: There is no significant difference in the effects of AI chatbots on learning outcomes at different educational levels at the 0.05 level.

Intervention duration

Previous studies have explored the effects of AI chatbots on learning outcomes in short duration. After taking 40 minutes to use AI chatbots, first-year undergraduate students showed more interest in learning (Yin et al., 2021). Nursing students, who used an AI chatbot-based learning approach in the two-week experiment, could acquire more knowledge than those without using AI chatbots (Chang et al., 2022). Researchers also tracked the effects of AI chatbots on learning outcomes over a long period of time. EFL learners improved their speaking skills after ten weeks of chatting with AI chatbots (Han, 2020). Through using AI chatbots for 16 weeks, there were significant improvements in students' English grammar skills (Kim, 2019). Therefore, we proposed the following hypothesis:

H8: There is no significant difference in the effects of AI chatbots on learning outcomes depending on the duration of AI chatbot use at the 0.05 level.

RESEARCH METHODS

Literature search

This meta-analysis was carried out based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page et al., 2021). On January 4, 2023, researchers conducted a wide-ranging and rigorous search of the literature in online academic databases, including Elsevier, Web of Science, Tailor & Francis, Wiley and Springer. The search process can be seen in Figure 1.

We obtained 141 results from Elsevier by keying in *Title, abstract, keywords*: "artificial intelligence chatbot" OR "artificial intelligence chatbot" OR "artificial intelligence-based chatbot" OR "AI chatbot" OR "AI-based chatbot" OR "artificial intelligence agent" OR "artificial intelligence agent" OR "AI agent" OR "AI conversational agent". We retrieved 369 results from Springer by entering in "chatbot" in "*where the title contains*" and "artificial intelligence" in "*with the exact phrase*". We collected 286 results from Taylor & Francis by keying in "chatbot OR agent" AND "artificial intelligence" in *abstract*. We obtained 105 results from Wiley by entering in "artificial intelligence" in *abstract* AND "chatbot* OR "agent*" in *title*. We got 486 results from Web of Science by entering in ("artificial intelligen* chatbot*" OR "AI chatbot*" OR "AI-based chatbot*" OR "artificial intelligen* agent*" OR "artificial intelligen* conversational agent*" OR "AI agent" OR "AI-based agent" OR "AI-based conversational agent") in the column *topic*.

Study selection

Publications would be included in this meta-analysis if they (a) focused specifically on the effects of AI chatbots on learning outcomes in terms of learning performance, motivation, self-efficacy, perceived value of learning, anxiety and interest; (b) had to adopt either a randomized controlled or quasi-experimental design; (c) were full-texts available to collect adequate information about study design and statistic results (i.e. mean and standardized deviation of both the control group and the intervention group); (d) were well-designed; (e) were written in English. Publications would be excluded if they (a) could not provide

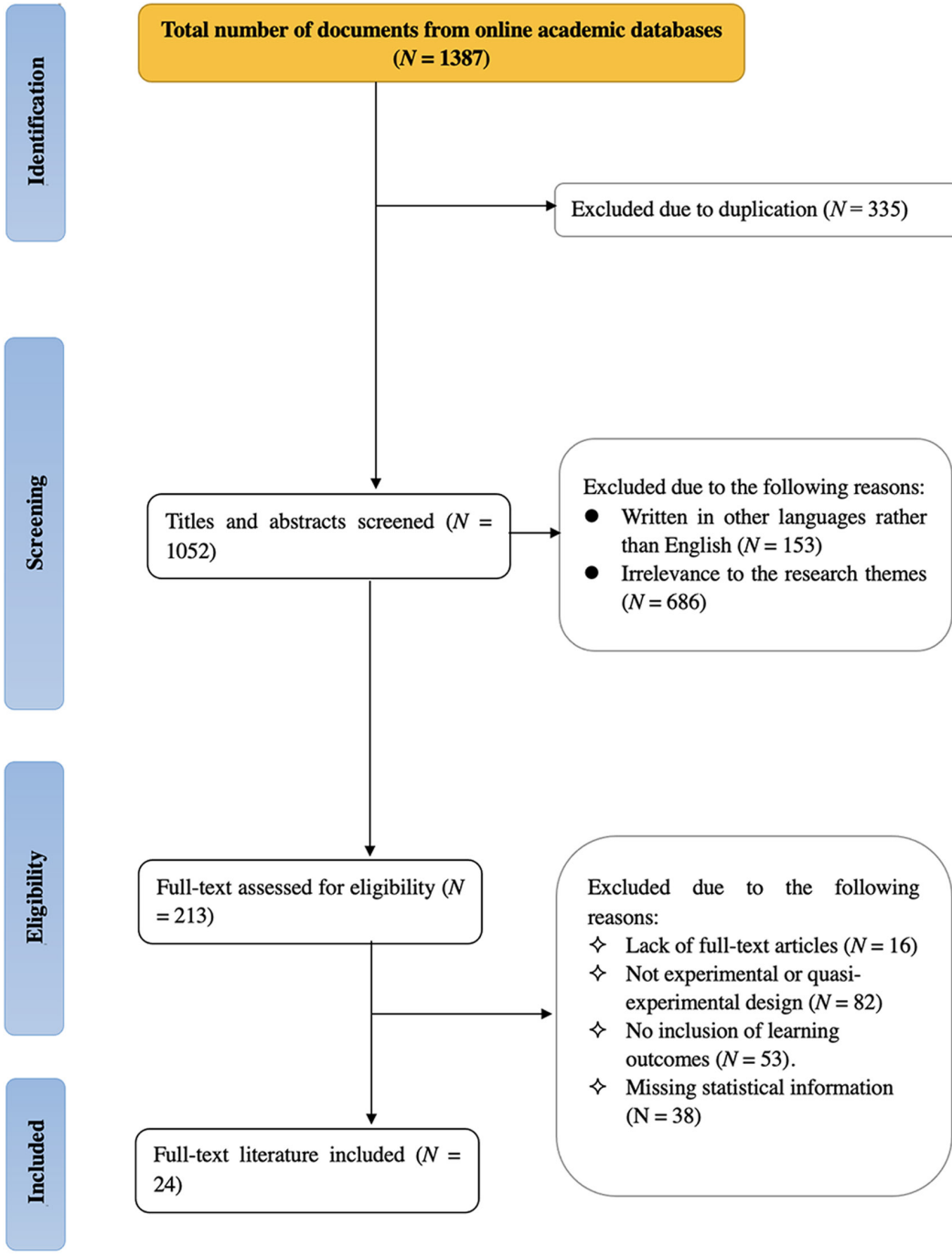

FIGURE 1 A flow chart of literature research.

adequate information for this meta-analysis even though we contacted the corresponding authors; (b) were written in other languages rather than English; (c) explored the AI chatbot itself instead of the effects of AI chatbot on learning outcomes; (d) used traditional chatbots rather than AI chatbots.

Researchers collected 24 studies that met the inclusion criteria. Two researchers independently carried out the quality assessment of each study using a quality assessment checklist validated by Kmet et al. (2004). The checklist comprised 14 items, which could be used to evaluate all sections of empirical studies, for example, study design, sample size and statistical analysis. The inter-rater reliability was 0.93. Disagreement on the quality assessment was discussed and resolved by two authors.

Study coding

Researchers followed the coding scheme proposed by Cooper (2016). Two researchers independently used the content analysis technique and retrieved the following key information: (a) author(s); (b) publication year; (c) sample size; (d) subgroups of learning outcomes; (e) educational levels; (f) intervention duration; (g) statistical results of learning outcomes (i.e. mean and standard deviation of both the control group and the intervention group). Educational levels were categorized as primary education, secondary education, and higher education. There has been disagreement on a cut-off point between long and short interventions (Ran et al., 2022). Many studies recommended the use of 10 weeks as a cut-off point to distinguish between long and short interventions (eg, Lin & Lin, 2019). Following the recommendation, this meta-analysis coded intervention duration as either short (shorter than 10 weeks) or long interventions (longer than 10 weeks). The inter-coder reliability was 0.94, showing a good agreement between the two researchers (Cohen, 1968). Disagreements on codes were discussed and resolved by two researchers. The Appendix A displays the characteristics of publications included in this meta-analysis.

Statistical analysis

We followed Yu's (2021) meta-analytic procedure. The meta-analytic procedure includes (1) the estimation of effect size, (2) the homogeneity analysis, (3) the moderator analysis, (4) publication bias tests and (5) the sensitivity analysis. All analyses were performed using Stata 14.0. Standardized mean differences (SMD) with a 95% confidence interval (CI) could be applied to estimate the effects of AI chatbots on learning outcomes. We calculated effect sizes as standardized mean differences using Cohen's *d* (Cohen, 1988). Positive Cohen's *d* values suggest that the chatbot group could have better learning outcomes than the control group (Villena-Taranilla et al., 2022). Additionally, the threshold values recommended by Cohen (1988) were used to interpret effect sizes (0.2 for small effect size, 0.5 for medium effect size, and 0.8 for large effect size).

Tests of homogeneity were performed to identify homogeneity across the included publications and to decide the type of model adopted in this meta-analysis (Hedges & Olkin, 1985). The potential homogeneity across the included publications could be assessed using Cochran's *Q* tests and I^2 statistics (Higgins & Thompson, 2002). Heterogeneity could be considered significant when the *p*-value is smaller than 0.05. The I^2 value is the percentage of homogeneity that might cause possible sources rather than sampling error (Higgins & Thompson, 2002; Theeboom et al., 2014). The heterogeneity will be important when the I^2 value ranges from 0% to 40%, moderate if the I^2 value ranges from 30% to 60%, substantial if the I^2 value ranges from 50% to 90%, and considerable if the I^2 value exceeds 75%. We will use a fixed-effect model if the I^2 value is lower than 50%. Otherwise, a random-effect model will be adequate for this meta-analysis (Higgins et al., 2011).

We proceeded to perform moderator analyses to identify whether factors could moderate the effects of AI chatbots on learning outcomes. The majority of previous meta-analyses found that intervention duration and educational levels could significantly moderate the effects of educational technologies on learning outcomes (eg, Bai et al., 2020; Theeboom et al., 2014; Wu, Yu et al., 2020). We thus hypothesized that intervention duration and educational levels might act as possible moderators in this meta-analysis. In the end, researchers carried out publication bias tests and sensitivity analysis.

RESULTS

Learning outcomes

As shown in Table 1, effect sizes were found statistically heterogeneous in terms of learning performance ($Q=372.22$, $df=20$, $I^2=94.6\%$, $p<0.001$), motivation ($Q=7.88$, $df=2$, $I^2=74.6\%$, $p=0.019$), self-efficacy ($Q=22.74$, $df=3$, $I^2=86.8\%$, $p<0.001$), interest ($Q=102.45$, $df=6$, $I^2=94.1\%$, $p=0.014$), anxiety ($Q=7.17$, $df=2$, $I^2=72.1\%$, $p=0.028$), perceived value of learning ($Q=57$, $df=3$, $I^2=94.7\%$, $p<0.001$), and overall learning outcomes ($Q=686.59$, $df=41$, $I^2=94\%$, $p<0.001$). Consequently, we adopted a random-effect model to calculate effect sizes.

Table 1 reveals a statistically significant large effect of AI chatbots on overall learning outcomes (ES = 0.964, 95% CI = [0.642, 1.286]). Specifically, using AI chatbots could significantly improve learning outcomes in terms of learning performance (ES = 1.028, 95% CI = [0.580, 1.476]), motivation (ES = 1.020, 95% CI = [0.278, 1.763]), self-efficacy (ES = 1.206, 95% CI = [0.357, 2.055]), interest (ES = 1.084, 95% CI = [0.220, 1.947]), and perceived value of learning (ES = 1.397, 95% CI = [0.228, 2.566]). In addition, using AI chatbots could significantly relieve learners' anxiety (ES = −0.715, 95% CI = [−1.302, −0.127]). Therefore, we accepted H1, H2, H3, H4, H5 and H6.

Educational levels

Table 2 shows that the overall effect sizes of AI chatbots on students at different educational levels could be considered significantly heterogeneous ($Q=686.59$, $df=41$, $I^2=94\%$, $p<0.001$). Specifically, effect sizes could be found significantly heterogeneous in primary education ($Q=18.51$, $df=2$, $I^2=89.2\%$, $p<0.001$), secondary education ($Q=33.66$, $df=4$, $I^2=88.1\%$, $p<0.001$), and higher education ($Q=625.89$, $df=33$, $I^2=94.7\%$, $p<0.001$). Accordingly, a random-effect model was used to test effect sizes at different educational levels.

As shown in Table 2, students in higher education obtained a large and significant effect size (ES = 1.079, 95% CI = [0.710, 1.448]). In contrast, effect sizes for primary school students (ES = 0.931, 95% CI = [−0.054, 1.916]) and secondary school students (ES = 0.214, 95% CI = [−0.608, 1.036]) are statistically insignificant. Therefore, there is a significant difference in the effect size of primary education, secondary education and higher education. It indicates that the effects of AI chatbots may vary according to educational levels. Therefore, we rejected H7.

Intervention duration

Tests of heterogeneity display that effect sizes were considered heterogeneous in different intervention duration (Table 2) in terms of less than ten weeks ($Q=414.65$, $df=26$, $I^2=93.7\%$,

TABLE 1 Meta-analysis results of the effects of AI chatbots.

Outcome	SMD	95%CI		%Weight	Cochran's *Q*	*df*	*p*	I^2 (%)	*z*	*p*
		LL	UL							
Learning performance	1.028	0.580	1.476	50.05	372.22	20	<0.001	94.6	4.49	<0.001
Learning motivation	1.020	0.278	1.763	7.02	7.88	2	0.019	74.6	2.69	0.007
Learning self-efficacy	1.206	0.357	2.055	9.33	22.74	3	<0.001	86.8	2.78	0.005
Learning interest	1.084	0.220	1.947	16.75	102.45	6	<0.001	94.1	2.45	0.014
Perceived value of learning	1.397	0.228	2.566	9.63	57	3	<0.001	94.7	2.34	0.019
Learning anxiety	−0.715	−1.302	−0.127	7.22	7.17	2	0.028	72.1	−2.38	0.017
Overall	0.964	0.642	1.286	100	686.59	41	<0.001	93.9	5.87	<0.001

Abbreviations: CI, confidence interval; *df*, degree of freedom; LL, low limit; SMD, standardized mean differences; UL, upper limit.

TABLE 2 Meta-analysis results of moderator analyses.

Moderator variable	SMD	95%CI		%Weight	Cochran's *Q*	*df*	*p*	I^2 (%)	*z*	*p*
		LL	UL							
Educational level										
Primary	0.931	−0.054	1.916	7.19	18.51	2	<0.001	89.2	1.85	0.064
Secondary	0.214	−0.608	1.036	11.75	33.66	4	<0.001	88.1	0.51	0.610
Higher	1.079	0.710	1.448	81.06	625.89	33	<0.001	94.7	5.73	<0.001
Overall	0.964	0.642	1.286	100	686.59	41	<0.001	94.0	5.87	<0.001
Duration										
<10 week	1.179	0.752	1.606	65.58	414.65	26	<0.001	93.7	5.41	<0.001
≥10 weeks	0.492	0.046	0.937	34.42	167.89	13	<0.001	92.3	2.16	0.031
Overall	0.945	0.620	1.269	100	660.32	40	<0.001	93.9	5.69	<0.001

Abbreviations: CI, confidence interval; *df*, degree of freedom; LL, low limit; SMD, standardized mean differences; UL, upper limit.

$p < 0.001$) and ten weeks or longer ($Q = 167.89$, $df = 13$, $I^2 = 92.3\%$, $p = 0.031$). In the same vein, the overall results present heterogeneous effect sizes ($Q = 660.32$, $df = 40$, $I^2 = 93.9\%$, $p < 0.001$). Thus, we adopted a random-effect model to test effect sizes in different intervention duration.

Table 2 shows that short interventions with a duration of shorter than ten weeks have a large and significant effect size (ES = 1.179, 95% CI = [0.752, 1.606]), whereas long interventions of a duration of ten weeks or longer yield a small and significant effect (ES = 0.492, 95% CI = [0.046, 0.937]). In other words, both short interventions and long interventions could be conducive to improving learning outcomes. It was noteworthy that the short interventions with a duration of shorter than ten weeks were more effective than the long interventions with a duration of ten weeks or longer. Therefore, we rejected H8.

Sensitivity analysis

In order to test the robustness of the meta-analysis results, we performed a sensitivity analysis using one by one elimination method (Figure 2). The meta-analysis results are not influenced when a given study was excluded. All estimates could fall within the area between the lower and upper 95% confidence limits (95% CI = [0.54, 1.29]).

Publication bias

Researchers evaluated the publication bias regarding learning outcomes using the funnel plot, Begg's test, Egger's test, and trim-and-fill test. From Figure 3, it is apparent that the funnel plot is slightly asymmetrical. Then, Egger's test and Begg's test were used to assess publication bias. Egger's test indicates an absence of publication bias in learning motivation

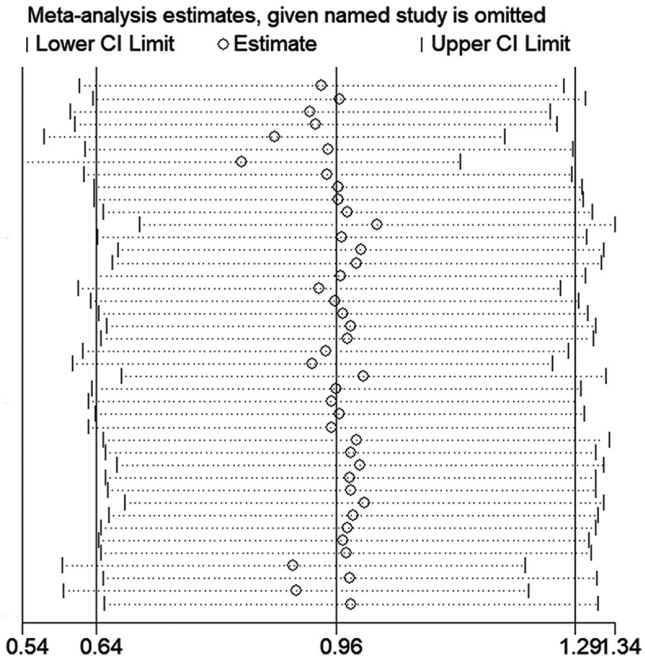

FIGURE 2 A plot of sensitivity analysis. CI, confidence interval.

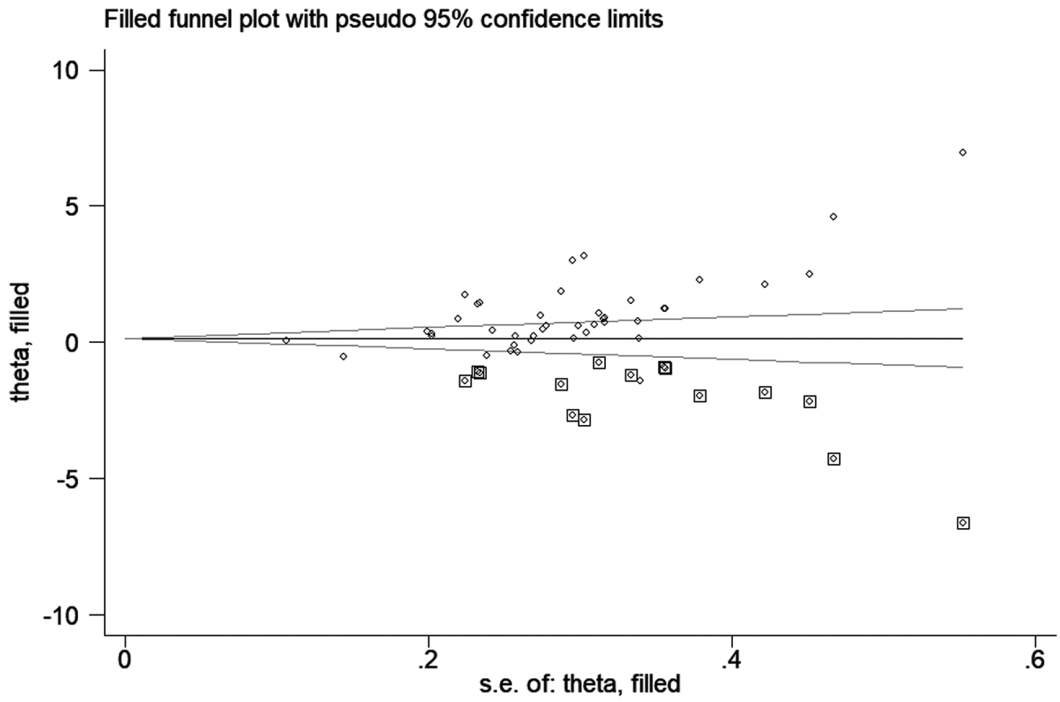

FIGURE 3 Funnel plot showing publication bias (Duval and Tweedie's trim-and-fill technique). Included studies are dots. Imputed studies are dots in the squares.

($z = 1.35$, $p = 0.405$), interest ($z = 0.89$, $p = 0.413$), perceived value of learning ($z = 0.49$, $p = 0.673$) and anxiety ($z = -2.79$, $p = 0.219$) but indicates a presence of publication bias in the overall outcomes ($z = 4.37$, $p < 0.001$), learning performance ($z = 4.13$, $p = 0.001$) and self-efficacy ($z = 6.86$, $p = 0.021$). Begg's test indicates an absence of publication bias in learning motivation ($z = 0.52$, $p = 0.602$), perceived value of learning ($z = -0.34$, $p = 0.872$), anxiety ($z = -0.52$, $p = 0.602$), self-efficacy ($z = 1.70$, $p = 0.089$), and interest ($z = 1.2$, $p = 0.230$) but a presence of publication bias in the overall outcomes ($z = 3.84$, $p = <0.001$) and learning performance ($z = 2.81$, $p = 0.005$).

The publication bias resulting from educational levels was measured using both Begg's and Egger's tests (Table 3). Egger's test indicates an absence of publication bias in terms of primary education ($z = 0.46$, $p = 0.724$) and secondary education ($z = -2.34$, $p = 0.101$) but indicates a presence of publication bias in terms of higher education ($z = 4.84$, $p < 0.001$) and overall educational levels ($z = 4.37$, $p < 0.001$). Begg's test indicates an absence of publication bias in terms of primary education ($z = 0.52$, $p = 0.602$) and secondary education ($z = 0.73$, $p = 0.462$) but a presence of publication bias in terms of higher education ($z = 4.18$, $p < 0.001$) and overall educational levels ($z = 3.84$, $p = <0.001$).

The publication bias stemming from intervention duration was also calculated using both Begg's test and Egger's test (see Table 3). Egger's test indicates an absence of publication bias regarding a duration of ten weeks or longer ($z = 2.02$, $p = 0.066$) but a presence of publication regarding a duration of shorter than ten weeks ($z = 3.78$, $p = 0.001$) and overall intervention duration ($z = 4.54$, $p = <0.001$). Begg's test indicates an absence of publication bias regarding the duration of ten weeks or longer ($z = 1.86$, $p = 0.063$), but a presence of publication regarding the duration of shorter than ten weeks ($z = 3.54$, $p = <0.001$) and overall intervention duration ($z = 4.14$, $p < 0.001$).

TABLE 3 Publication bias.

N	Variable	Egger's				Begg's	
		Score	*SD*	*z*	*p*	*z*	*p*
Learning outcomes							
1	Learning performance	7.567	1.834	4.13	0.001	2.81	0.005
2	Learning motivation	19.779	14.605	1.35	0.405	0.52	0.602
3	Learning self-efficacy	7.927	1.155	6.86	0.021	1.70	0.089
4	Learning interest	10.501	11.777	0.89	0.413	1.20	0.230
5	Perceived value of learning	9.925	20.253	0.49	0.673	−0.34	0.872
6	Anxiety	−10.074	3.606	−2.79	0.219	−0.52	0.602
	Overall	7.11	1.626	4.37	<0.001	3.84	<0.001
Educational level							
1	Primary	12.308	26.596	0.46	0.724	0.52	0.602
2	Secondary	−53.46	23.02	−2.34	0.101	0.73	0.462
3	Higher	8.209	1.696	4.84	<0.001	4.18	<0.001
4	Overall	7.11	1.626	4.37	<0.001	3.84	<0.001
Duration							
1	<10 week	10.892	2.881	3.78	0.001	3.54	<0.001
2	≥10 weeks	4.281	2.118	2.02	0.066	1.86	0.063
3	Overall	7.213	1.589	4.54	<0.001	4.14	<0.001

Abbreviations: N, number; *SD*, standard deviation.

In addition, Duval and Tweedie's trim-and-fill technique was employed to test publication bias and obtain an unbiased estimate of the effect size (Duval & Tweedie, 2000). The trim-and-fill technique recommended that 15 hypothetical missing studies should be added in order to obtain a symmetric funnel plot (Figure 3). It indicates a presence of publication bias, possibly due to the influence of small study bias on learning outcomes. The imputed effect size was 1.212 (95% CI = [0.841, 1.745]), indicating a significantly large effect of AI chatbots on overall learning outcomes ($p < 0.001$).

DISCUSSION

Due to the widespread popularity of AI chatbots, many studies have devoted much attention to the effects of using AI chatbots on learning outcomes. Prior empirical research provided mixed evidence for the effects of AI chatbots on learning outcomes. This study, through a meta-analysis, synthesized the inconsistent results and explored the effects of AI chatbots on students' learning outcomes. Additionally, this meta-analysis explored the moderating effects of educational levels and intervention duration.

Students assisted by AI chatbots could achieve better learning performance. This finding further supports the idea that AI chatbots could enhance learning performance (Kim, 2019; Mageira et al., 2022). The affordances of high personalization and interactivity in AI chatbots may bring increased learning performance. AI chatbots could provide students with one-on-one tutoring and needed exercises so that students could solidify their understanding of

knowledge (Hwang et al., 2022). Besides, AI chatbots could provide diagnostic information about individual students' performance in the learning process (Kim et al., 2022). The diagnostic information could help students focus on individual problems in the learning process, subsequently improving competence satisfaction and learning performance (Jeon, 2021). During the review process, AI chatbots could help learners integrate new knowledge and old knowledge through deeper discussions and efficient interactions (Lee et al., 2022). Therefore, these affordances of AI chatbots may explain why using AI chatbots had a stronger effect on students' learning performance.

AI chatbots could lead to significantly higher learning motivation among students. This result could echo the findings of Jeon (2022), which showed that AI chatbots improved students' learning motivation. There are several possible reasons for this result. Many AI chatbots could be implemented in various digital devices such as smartphones and laptops, enabling students to study anywhere and improve learning motivation (Huang et al., 2022). Providing detailed feedback in the learning process, AI chatbots were effective in lowering students' extraneous cognitive load and enhancing their learning motivation (Fidan & Gencel, 2022). Human-like AI chatbots could stimulate students' learning motivation, possibly due to their delivery of adequate attention, emotional exchange, and a sense of social presence (Ebadi & Amini, 2022).

AI chatbots could significantly enhance learning self-efficacy. The result is in keeping with that of Fryer et al. (2020). AI chatbots could help students promptly deal with learning problems, promoting their problem-solving abilities and self-efficacy in learning (Chang et al., 2022). AI chatbots could offer more training opportunities so that students developed learning self-efficacy through continuous imitation and practice (Hsu et al., 2021). AI chatbots could help students to improve learning self-efficacy by expressing their agreement with students' ideas during the conversation (Huang et al., 2022).

AI chatbots could greatly inspire and sustain students' learning interests. This finding corroborates the idea of Haristiani (2019) who suggested that AI chatbots significantly aroused learning interests. AI chatbots could allow students to choose preferred learning styles (Yin et al., 2021) and provide visible learning cues to assist students to finish learning tasks (Liu, Liao et al., 2022). Moreover, learning and interacting with AI chatbots may be easy and fun for students. Therefore, it was reasonable to find that AI chatbots could inspire and sustain learners' learning interests.

The use of AI chatbots could lead to an increased perceived value of learning among students. The result corroborates the idea of Fidan and Gencel (2022), who suggested the positive effects of AI chatbots on students' perceived value of learning. AI chatbots could provide learners with more flexible options to learn and practice, delivering more value to their understanding of knowledge (Yin et al., 2021). AI chatbots could help students to be immersed in the learning environment, increasing students' perceived value of reading (Liu, Liao et al., 2022).

AI chatbots could be helpful to alleviate learners' anxiety. The result supports a recent study, which found that AI chatbots reduced EFL learners' speaking anxiety (Han, 2020). AI chatbots could create a relaxed self-learning environment for students who felt shy and anxious about making mistakes in front of peers and teachers (Hsu et al., 2021; Jeon, 2022). Besides, many AI chatbots could detect students' emotions and then provide emotional support, guidance and advice based on students' expressions so that students could alleviate anxiety and stress in the learning process (Klos et al., 2021).

The effects of AI chatbots were different across educational levels. University students' learning outcomes significantly improved from the support of AI chatbots. However, this finding is not observed among primary school and secondary school students. These results corroborate the idea of Garzón and Acevedo (2019), who also found that educational technologies had a greater effect on university students than on primary school

and secondary school students. Compared to other age groups, primary school students may not always have an effective interaction with AI chatbots due to lower language competency, self-directed learning ability and digital literacy (Jeon, 2022). They thus relied more on teachers' guidance, which made it difficult for them to fully engage with learning through AI chatbots (Deveci Topal et al., 2021). Secondary school students received more academic pressure from their parents and schools for good exam results (Tang et al., 2020). High academic pressure decreased their motivation to use AI chatbots for learning. By contrast, university students may be more proactive and capable of self-regulating their learning (Xu et al., 2022). It was hardly surprising that they had a higher level of engagement and better learning outcomes in chatbot-based learning, compared to other age groups.

The duration of AI chatbot use had a significant effect on students' learning outcomes. The short interventions had a greater effect on students' learning outcomes than the long interventions. This result is in accord with a recent study finding that learning outcomes were greatly improved by education technologies during short-term interventions rather than long-term interventions (Villena-Taranilla et al., 2022). The result may be due to the novelty effect of AI chatbots. A growing body of evidence shows that the novelty effect of educational technologies could improve students' interest, motivation and engagement. Nevertheless, it may wear off once students have used those technologies for a long time (Jeno et al., 2019). It was reasonable to infer that in the short interventions, students were excited and intrigued by the use of AI chatbots, which temporarily increased their learning interest, motivation, and performance (Fryer et al., 2017). Once the novelty effect of AI chatbots has worn off, it would probably be difficult for students to sustain learning interests, have motivation, and perform well in chatbot-based learning (Haristiani, 2019). Additionally, students were more likely to experience information overload when using AI chatbots for a long time. This may distract them from chatbot-supported learning, leading to poor learning performance. The duration of chatbot use is an important issue that future studies should focus on.

CONCLUSION

Major findings

This meta-analysis set out to integrate the findings of multiple empirical studies to identify the effects of AI chatbots on learning outcomes. Our results indicated that AI chatbots could have a large effect on students' learning outcomes in terms of performance, motivation, interest, self-efficacy, perceived value of learning and anxiety. Additionally, this meta-analysis found that educational levels and intervention duration could moderate the effects of AI chatbots on learning outcomes. The results indicated that students in higher education seemed to benefit the most from AI chatbots. By contrast, primary and secondary school students assisted with AI chatbots could not have better learning outcomes than those without using AI chatbots. Regarding the intervention duration, short interventions were found to be more effective in enhancing students' learning outcomes, compared to long interventions.

Major contributions

This study makes an important contribution to research on AI chatbots by demonstrating that AI chatbots have a large effect on students' learning performance, motivation, interest, self-efficacy, anxiety and perceived value of learning. Besides, this study identified educational

levels and durations that may moderate the effects of AI chatbots on learning outcomes. The moderator analyses could provide valuable insight into how to effectively make use of AI chatbots for learning. Finally, this study presents a further step towards developing AI chatbots and proposes recommendations for future work, which could develop a better understanding of using AI chatbots in education.

Limitations

This meta-analysis is subjected to some limitations. First, this study had to exclude some empirical studies (eg, Deveci Topal et al., 2021) because they did not provide insufficient statistical data for the calculation of effect sizes (eg, means and standard deviations of both control and experimental groups). It could potentially modify the meta-analysis results. Second, there was some publication bias in this meta-analysis since studies with negative results were more likely to be rejected than those with positive ones. Third, we left out some related publications due to the inaccessible databases. Fourth, because of the limitations in the statistical data provided by the included publications, we could not explore the effects of other moderators, such as the context of chatbot use, learning locations, gender, culture, types of AI chatbots and learning activities. Given the above limitations, the findings about the effects of AI chatbots on learning outcomes need to be interpreted with caution.

Suggestions

Suggestions for AI chatbot designers

There are some suggestions for AI chatbot designers. As this study demonstrated the effects of AI chatbots on learning engagement, AI chatbot designers could foster student-chatbot interactions and learning engagement by equipping AI chatbots with human-like avatars. Given the important roles of AI chatbots in relieving learning anxiety, AI chatbot designers could equip AI chatbots with emotional intelligence so that they could recognize students' emotional states, express empathy and provide affective guidance. They might be more useful for students to alleviate learning anxiety. Our findings also call on AI chatbot designers to add gamification elements to AI chatbots, which could be effective to enhance students' learning performance, interest and motivation.

Suggestions for future studies

Future research should take into account both learning outcomes and negative effects. Numerous studies highlighted the positive effects of using AI chatbots in education. However, some studies found that users felt frustrated and lost learning motivation when using AI chatbots to practice communication skills (Kim et al., 2021). Relatively little research has reported and analysed the negative effects of AI chatbots on students' learning outcomes. More research is needed to examine both the positive and negative effects of using AI chatbots in education. It will be useful to provide a more comprehensive evaluation of the potential for using AI chatbots in education.

ChatGPT is an emerging AI chatbot developed by OpenAI. Over the past few weeks, ChatGPT has been attracting considerable attention because of its advanced language processing capacity (Pavlik, 2023). Specifically, ChatGPT is capable of understanding natural language inputs and generating remarkably intelligent responses within seconds

(Rudolph et al., 2023). A growing body of literature indicates that there is a lot of potential for using ChatPGT in education. For example, ChatGPT could grade written assignments and provide constructive suggestions for improving writing and speaking (van Dis et al., 2023). Additionally, ChatGPT could provide personalized learning for a large number of students anytime and anywhere (Mogali, 2023). Nevertheless, there is a notable paucity of empirical research focusing specifically on the use of ChatGPT in education. More empirical research on ChatGPT needs to be undertaken.

The moderating analyses could provide a more comprehensive understanding of the effects of AI chatbots. It is thus meaningful to investigate whether some variables could moderate the effects of AI chatbots on students' learning outcomes. For example, future studies could examine whether there are significant gender and cultural differences in AI chatbot-supported learning outcomes. Moreover, students with different personality traits could have different learning styles (Kamal & Radhakrishnan, 2019). Future research could explore what type of learners benefit most from AI chatbot-supported learning. Furthermore, it would be interesting to compare which type of AI chatbots may work best in education. It is important to note that students' learning outcomes might be influenced by the context of AI chatbot use. Further research is needed to better understand the moderating role of the context of AI chatbot use.

ACKNOWLEDGEMENTS

This work is supported by 2019 MOOC of Beijing Language and Culture University (MOOC201902) (Important) “Introduction to Linguistics”; “Introduction to Linguistics” of online and offline mixed courses in Beijing Language and Culture University in 2020; Special fund of Beijing Co-construction Project-Research and reform of the “Undergraduate Teaching Reform and Innovation Project” of Beijing higher education in 2020-innovative “multilingual +” excellent talent training system (202010032003); The research project of Graduate Students of Beijing Language and Culture University “Xi Jinping: The Governance of China” (SJTS202108); the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (22YCX038); Beijing Language and Culture University Excellent Doctoral Dissertation Cultivation Program Funding Project.

CONFLICT OF INTEREST STATEMENT

We have no conflicts of interest to declare that are relevant to the content of this article.

DATA AVAILABILITY STATEMENT

We make sure that all data and materials support our published claims and comply with field standards.

ETHICS STATEMENT

The study was approved by the institutional review board of Beijing Language and Culture University. All researchers can provide written informed consents.

ORCID

Rong Wu https://orcid.org/0000-0002-2985-1197
Zhonggen Yu https://orcid.org/0000-0002-3873-980X

REFERENCES

*Abbasi, S., Kazi, H., & Hussaini, N. (2019). Effect of chatbot systems on student's learning outcomes. *Sylwan*, *163*(10), 49–63.

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, Article 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review*, *30*, Article 100322. https://doi.org/10.1016/j.edurev.2020.100322

Brachten, F., Brünker, F., Frick, N. R. J., Ross, B., & Stieglitz, S. (2020). On the ability of virtual agents to decrease cognitive load: An experimental study. *Information Systems and e-Business Management*, *18*(2), 187–207. https://doi.org/10.1007/s10257-020-00471-7

*Chang, C. Y., Hwang, G. J., & Gau, M. L. (2022). Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training. *British Journal of Educational Technology*, *53*(1), 171–188. https://doi.org/10.1111/bjet.13158

Chien, Y. C., Wu, T. T., Lai, C. H., & Huang, Y. M. (2022). Investigation of the influence of artificial intelligence markup language-based LINE chatbot in contextual English learning. *Frontiers in Psychology*, *13*, Article 785752. https://doi.org/10.3389/fpsyg.2022.785752

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220. https://doi.org/10.1037/h0026256

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach*. SAGE.

Deveci Topal, A., Dilek Eren, C., & Kolburan Geçer, A. (2021). Chatbot application in a 5th grade science course. *Education and Information Technologies*, *26*(5), 6241–6265. https://doi.org/10.1007/s10639-021-10627-8

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Ebadi, S., & Amini, A. (2022). Examining the roles of social presence and human-likeness on Iranian EFL learners' motivation using artificial intelligence technology: A case of CSIEC chatbot. *Interactive Learning Environments*, 1–19. https://doi.org/10.1080/10494820.2022.2096638

*Essel, H. B., Vlachopoulos, D., Tachie-Menson, A., Johnson, E. E., & Baah, P. K. (2022). The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. *International Journal of Educational Technology in Higher Education*, *19*(1), Article 57. https://doi.org/10.1186/s41239-022-00362-6

*Fidan, M., & Gencel, N. (2022). Supporting the instructional videos with chatbot and peer feedback mechanisms in online learning: The effects on learning performance and intrinsic motivation. *Journal of Educational Computing Research*, *60*(7), 1716–1741. https://doi.org/10.1177/07356331221077901

Fryer, L., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning & Technology*, *24*(2), 8–22. http://hdl.handle.net/10125/44719

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners. *Computers in Human Behavior*, *75*, 461–468. https://doi.org/10.1016/j.chb.2017.05.045

Garrido, D., Martín, M. V., Rodríguez, C., Iglesias, J., Navarro, J. C., Estévez, A., Hontoria, F., Becerro, M., Otero, J. J., Pérez, J., Varó, I., Reis, D. B., Riera, R., Sykes, A. V., & Almansa, E. (2018). Meta-analysis approach to the effects of live prey on the growth of Octopus vulgaris paralarvae under culture conditions. *Reviews in Aquaculture*, *10*(1), 3–14. https://doi.org/10.1111/raq.12142

Garzón, J., & Acevedo, J. (2019). Meta-analysis of the impact of augmented reality on students' learning gains. *Educational Research Review*, *27*, 244–260. https://doi.org/10.1016/j.edurev.2019.04.001

*Han, D. E. (2020). The effects of voice-based AI chatbots on Korean EFL middle school students' speaking competence and affective domains. *Asia-Pacific Journal of Convergent Research Interchange*, *6*(7), 71–80. https://doi.org/10.47116/apjcri.2020.07.07

*Han, J. W., Park, J., & Lee, H. (2022). Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: A quasi-experimental study. *BMC Medical Education*, *22*(1), Article 830. https://doi.org/10.1186/s12909-022-03898-3

Haristiani, N. (2019). Artificial intelligence (AI) chatbot as language learning medium: An inquiry. *Journal of Physics: Conference Series*, *1387*(1), Article 012020. https://doi.org/10.1088/1742-6596/1387/1/012020

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L., & Sterne, J. A. C. (2011). The cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, *343*, Article 5928. https://doi.org/10.1136/bmj.d5928

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hsu, M. H., Chen, P. S., & Yu, C. S. (2021). Proposing a task-oriented chatbot system for EFL learners speaking practice. *Interactive Learning Environments*, 1–12. https://doi.org/10.1080/10494820.2021.1960864

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257. https://doi.org/10.1111/jcal.12610

*Hwang, W. Y., Guo, B. C., Hoang, A., Chang, C. C., & Wu, N. T. (2022). Facilitating authentic contextual EFL speaking and conversation with smart mechanisms and investigating its influence on learning achievements. *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2022.2095406

Jeno, L. M., Vandvik, V., Eliassen, S., & Grytnes, J. A. (2019). Testing the novelty effect of an m-learning tool on internalization and achievement: A self-determination theory approach. *Computers & Education*, *128*, 398–413. https://doi.org/10.1016/j.compedu.2018.10.008

*Jeon, J. (2021). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2021.1987272

Jeon, J. (2022). Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning*, 1–26. https://doi.org/10.1080/09588221.2021.2021241

Kamal, A., & Radhakrishnan, S. (2019). Individual learning preferences based on personality traits in an E-learning scenario. *Education and Information Technologies*, *24*(1), 407–435. https://doi.org/10.1007/s10639-018-9777-4

Kim, H. S., Kim, N. Y., & Cha, Y. J. (2021). Effects of AI chatbots on EFL students' communication skills. *Korean Journal of English Language and Linguistics*, *21*, 712–734. https://doi.org/10.15738/KJELL.21.202108.712

Kim, J. H., Kim, M., Kwak, D. W., & Lee, S. (2022). Home-tutoring services assisted with technology: Investigating the role of artificial intelligence using a randomized field experiment. *Journal of Marketing Research*, *59*(1), 79–96. https://doi.org/10.1177/00222437211050351

*Kim, N. Y. (2018a). A study on chatbots for developing Korean college students' English listening and reading skills. *Journal of Digital Convergence*, *16*(8), 19–26. https://doi.org/10.14400/JDC.2018.16.8.019

*Kim, N. Y. (2018b). Chatbots and Korean EFL students' English vocabulary learning. *Journal of Digital Convergence*, *16*(2), 1–7. https://doi.org/10.14400/JDC.2018.16.2.001

*Kim, N. Y. (2019). A study on the use of artificial intelligence chatbots for improving English grammar skills. *Journal of Digital Convergence*, *17*(8), 37–46. https://doi.org/10.14400/JDC.2019.17.8.037

*Klos, M. C., Escoredo, M., Joerin, A., Lemos, V. N., Rauws, M., & Bunge, E. L. (2021). Artificial intelligence–based chatbot for anxiety and depression in university students: Pilot randomized controlled trial. *JMIR Formative Research*, *5*(8), Article 20678. https://doi.org/10.2196/20678

Kmet, L. M., Lee, R. C., & Cook, L. S. (2004). *Standard quality assessment criteria for evaluating primary research papers from a variety of fields*. Alberta Heritage Foundation for Medical Research. https://www.deslibris.ca/ID/200548

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2022). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

*Kumar, J. A. (2021). Educational chatbots for project-based learning: Investigating learning outcomes for a team-based design course *International Journal of Educational Technology in Higher Education*, *18*(1), Article 65. https://doi.org/10.1186/s41239-021-00302-w

Kuo, T. M., Tsai, C. C., & Wang, J. C. (2021). Linking web-based learning self-efficacy and learning engagement in MOOCs: The role of online academic hardiness. *The Internet and Higher Education*, *51*, Article 100819. https://doi.org/10.1016/j.iheduc.2021.100819

*Lee, Y. F., Hwang, G. J., & Chen, P. Y. (2022). Impacts of an AI-based chatbot on college students' after-class review, academic performance, self-efficacy, learning attitude, and motivation. *Educational Technology Research and Development*, *70*(5), 1843–1865. https://doi.org/10.1007/s11423-022-10142-8

Lin, C. J., & Mubarok, H. (2021). Learning analytics for investigating the mind map-guided AI chatbot approach in an EFL flipped speaking classroom. *Educational Technology & Society*, *24*(4), 16–35.

Lin, J. J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*, *32*(8), 878–919. https://doi.org/10.1080/09588221.2018.1541359

*Lin, M. P. C., & Chang, D. (2020). Enhancing post-secondary writers' writing skills with a chatbot: A mixed-method classroom study. *Journal of Educational Technology & Society*, *23*(1), 78–92.

*Liu, C. C., Liao, M. G., Chang, C. H., & Lin, H. M. (2022). An analysis of children' interaction with an AI chatbot and its impact on their interest in reading. *Computers & Education*, *189*, Article 104576. https://doi.org/10.1016/j.compedu.2022.104576

*Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interventions*, *27*, Article 100495. https://doi.org/10.1016/j.invent.2022.100495

Luo, B., Lau, R. Y. K., Li, C., & Si, Y. W. (2022). A critical review of state-of-the-art chatbot designs and applications. *WIREs Data Mining and Knowledge Discovery*, *12*(1), Article 1434. https://doi.org/10.1002/widm.1434

*Mageira, K., Pittou, D., Papasalouros, A., Kotis, K., Zangogianni, P., & Daradoumis, A. (2022). Educational AI chatbots for content and language integrated learning. *Applied Sciences*, *12*(7), Article 7. https://doi.org/10.3390/app12073239

Mogali, S. R. (2023). Initial impressions of ChatGPT for anatomy education. *Anatomical Sciences Education*, 1–4. https://doi.org/10.1002/ase.2261

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital Health*, *5*, 1–12. https://doi.org/10.1177/2055207619871808

*Nghi, T., Huu Phuc, T., & Nguyen Tat, T. (2019). Applying AI chatbot for teaching a foreign language: An empirical research. *International Journal of Scientific & Technology Research*, *8*(12), 897–902.

Nguyen, Q. N., Sidorova, A., & Torres, R. (2022). User interactions with chatbot interfaces vs. menu-based interfaces: An empirical study. *Computers in Human Behavior*, *128*, Article 107093. https://doi.org/10.1016/j.chb.2021.107093

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), Article 1. https://doi.org/10.1186/s13643-021-01626-4

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, *78*(1), 84–93. https://doi.org/10.1177/10776958221149577

Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, *28*(6), 1549–1565. https://doi.org/10.1002/cae.22326

Pickering, C., & Byrne, J. (2014). The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers. *Higher Education Research & Development*, *33*(3), 534–548. https://doi.org/10.1080/07294360.2013.841651

Ran, H., Kim, N. J., & Secada, W. G. (2022). A meta-analysis on the effects of technology's functions and roles on students' mathematics achievement in K-12 classrooms. *Journal of Computer Assisted Learning*, *38*(1), 258–284. https://doi.org/10.1111/jcal.12611

*Ruan, S., Jiang, L., Xu, Q., Liu, Z., Davis, G. M., Brunskill, E., & Landay, J. A. (2021). EnglishBot: An AI-powered conversational system for second language learning. *In Proceedings of 26th International Conference on Intelligent User Interfaces* (pp. 434–444). Association for Computing Machinery https://doi.org/10.1145/3397481.3450648

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1), Article 1. https://doi.org/10.37074/jalt.2023.6.1.9

Salas-Pilco, S. Z. (2020). The impact of AI and robotics on physical, social-emotional and intellectual learning outcomes: An integrated analytical framework. *British Journal of Educational Technology*, *51*(5), 1808–1825. https://doi.org/10.1111/bjet.12984

Shumanov, M., & Johnson, L. (2021). Making conversations with chatbots more personalized. *Computers in Human Behavior*, *117*, Article 106627. https://doi.org/10.1016/j.chb.2020.106627

Tang, X., Tang, S., Ren, Z., & Wong, D. F. K. (2020). Psychosocial risk factors associated with depressive symptoms among adolescents in secondary schools in mainland China: A systematic review and meta-analysis. *Journal of Affective Disorders*, *263*, 155–165. https://doi.org/10.1016/j.jad.2019.11.118

Terblanche, N., Molyn, J., De Haan, E., & Nilsson, V. O. (2022). Coaching at scale: Investigating the efficacy of artificial intelligence coaching. *International Journal of Evidence Based Coaching and Mentoring*, *20*(2), 20–36. https://doi.org/10.24384/5cgf-ab69

Theeboom, T., Beersma, B., & van Vianen, A. E. M. (2014). Does coaching work? A meta-analysis on the effects of coaching on individual level outcomes in an organizational context. *The Journal of Positive Psychology*, *9*(1), 1–18. https://doi.org/10.1080/17439760.2013.837499

Tsai, Y., Lin, C., Hong, J., & Tai, K. (2018). The effects of metacognition on online learning interest and continuance to learn with MOOCs. *Computers & Education*, *121*, 18–29. https://doi.org/10.1016/j.compedu.2018.02.011

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

*Vázquez-Cano, E., Mengual-Andrés, S., & López-Meneses, E. (2021). Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments *International Journal of Educational Technology in Higher Education*, *18*(1), Article 33. https://doi.org/10.1186/s41239-021-00269-8

Villena-Taranilla, R., Tirado-Olivares, S., Cózar-Gutiérrez, R., & González-Calero, J. A. (2022). Effects of virtual reality on learning outcomes in K-6 education: A meta-analysis. *Educational Research Review*, *35*, Article 100434. https://doi.org/10.1016/j.edurev.2022.100434

*Wambsganss, T., Kueng, T., Soellner, M., & Leimeister, J. M. (2021). ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445781

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, *4*, Article 6454924. https://doi.org/10.3389/frai.2021.654924

Wu, B., Yu, X., & Gu, X. (2020). Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis. *British Journal of Educational Technology*, *51*(6), 1991–2005. https://doi.org/10.1111/bjet.13023

Wu, E. H. K., Lin, C. H., Ou, Y. Y., Liu, C. Z., Wang, W. K., & Chao, C. Y. (2020). Advantages and constraints of a hybrid model K-12 e-learning assistant chatbot. *IEEE Access*, *8*, 77788–77801. https://doi.org/10.1109/ACCESS.2020.2988252

Xu, W. W., Su, C. Y., Hu, Y., & Chen, C. H. (2022). Exploring the effectiveness and moderators of augmented reality on science learning: A meta-analysis. *Journal of Science Education and Technology*, *31*(5), 621–637. https://doi.org/10.1007/s10956-022-09982-z

*Yin, J., Goh, T. T., Yang, B., & Xiaobin, Y. (2021). Conversation technology with micro-learning: The impact of chatbot-based learning on students' learning motivation and performance. *Journal of Educational Computing Research*, *59*(1), 154–177. https://doi.org/10.1177/0735633120952067

Yu, Z. G. (2021). A meta-analysis of the effect of virtual reality technology use in education. *Interactive Learning Environments*, 1–21. https://doi.org/10.1080/10494820.2021.1989466

Zhang, J., Oh, Y. J., Lange, P., Yu, Z., & Fukuoka, Y. (2020). Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: Viewpoint. *Journal of Medical Internet Research*, *22*(9), Article 22845. https://doi.org/10.2196/22845

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine*, *5*(1), Article 1. https://doi.org/10.1038/s41746-022-00589-7

How to cite this article: Wu, R., & Yu, Z. (2024). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology*, *55*, 10–33. https://doi.org/10.1111/bjet.13334

APPENDIX A

The characteristics of publications included.

	Population characteristics				
	Sample size				
Authors (year)	EG	CG	Education level	Duration	Learning outcomes
Abbasi et al. (2019)	55	55	Higher education	*N*	Performance
Brachten et al. (2020)	45	46	Higher education	<10 weeks	Performance
Chang et al. (2022)	18	18	Higher education	<10 weeks	Performance, self-efficacy
Essel et al. (2022)	34	34	Higher education	≥10 weeks	Performance
Fidan and Gencel (2022)	54	40	Higher education	<10 weeks	Performance, interest, perceived value of learning
Han (2020)	22	22	Secondary education	≥10 weeks	Performance, interest, motivation, anxiety
Han et al. (2022)	30	31	Higher education	<10 weeks	Performance, interest
Hwang et al. (2022)	23	20	Primary education	≥10 weeks	Performance
Jeon (2021)	36	34	Primary education	<10 weeks	Performance
Kim (2018a)	24	22	Higher education	<10 weeks	Performance
Kim (2018b)	24	23	Higher education	≥10 weeks	Performance, motivation, interest, self-efficacy
Kim (2019)	36	34	Higher education	<10 weeks	Performance
Klos et al. (2021)	39	34	Higher education	<10 weeks	Anxiety
Kumar (2021)	30	30	Higher education	≥10 weeks	Performance
Lee et al. (2022)	18	20	Higher education	<10 weeks	Performance, motivation, self-efficacy
Lin and Chang (2020)	167	190	Higher education	≥10 weeks	Performance
Liu, Liao et al. (2022)	41	21	Primary education	<10 weeks	Perceived value of learning, Interest
Liu, Peng et al. (2022)	30	33	Higher education	≥10 weeks	Anxiety
Mageira et al. (2022)	18	17	Secondary education	<10 weeks	Performance
Nghi et al. (2019)	100	100	Higher education	≥10 weeks	Learning performance
Ruan et al. (2021)	28	28	Higher education	<10 weeks	Performance
Vázquez-Cano et al. (2021)	52	51	Higher education	≥10 weeks	Performance
Wambsganss et al. (2021)	31	24	Higher education	<10 weeks	Performance, interest
Yin et al. (2021)	51	48	Higher education	<10 weeks	Performance, interest, Perceived value of learning, self-efficacy

Abbreviations: CG, control group; EG, experimental group; N, no data reported.