Computers & Education 227 (2025) 105224

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies

Ruiqi Deng [a,b,*], Maoli Jiang [a], Xinlu Yu [a], Yuyan Lu [a], Shasha Liu [c]

[a] Jing Hengyi School of Education, Hangzhou Normal University, Hangzhou, China
[b] Chinese Education Modernization Research Institute (Zhejiang Provincial Key Think Tank), Hangzhou Normal University, Hangzhou, China
[c] School of Tourism Planning and Design, Tourism College of Zhejiang, Hangzhou, China

ARTICLE INFO

Keywords:
Teaching/learning strategies
Improve classroom teaching
Elementary education
Secondary education
Post-secondary education

ABSTRACT

Chat Generative Pre-Trained Transformer (ChatGPT) has generated excitement and concern in education. While cross-sectional studies have highlighted correlations between ChatGPT use and learning performance, they fall short of establishing causality. This review examines experimental studies on ChatGPT's impact on student learning to address this gap. A comprehensive search across five databases identified 69 articles published between 2022 and 2024 for analysis. The findings reveal that ChatGPT interventions are predominantly implemented at the university level, cover various subject areas focusing on language education, are integrated into classroom environments as part of regular educational practices, and primarily involve direct student use of ChatGPT. Overall, ChatGPT *improves* academic performance, affective-motivational states, and higher-order thinking propensities; it *reduces* mental effort and has *no* significant effect on self-efficacy. However, methodological limitations, such as the lack of power analysis and concerns regarding post-intervention assessments, warrant cautious interpretation of results. This review presents four propositions from the findings: (1) distinguish between the quality of ChatGPT outputs and the positive effects of interventions on academic performance by shifting from well-defined problems in post-intervention assessments to more complex, project-based assessments that require skill demonstration, adopting proctored assessments, or incorporating metrics such as originality alongside quality; (2) evaluate long-term impacts to determine whether the positive effects on affective-motivational states are sustained or merely owing to novelty effect; (3) prioritise objective measures to complement subjective assessments of higher-order thinking; and (4) use power analysis to determine adequate sample sizes to avoid Type II errors and provide reliable effect size estimates. This review provides valuable insights for researchers, instructors, and policymakers evaluating the effectiveness of generative AI integration in educational practice.

1. Introduction

Although large language models (LLMs) can be traced back to the early development of natural language processing in the mid-20th century (Maatouk et al., 2024), the release of Chat Generative Pre-Trained Transformer (ChatGPT) in late 2022 marked a turning point

* Corresponding author. Jing Hengyi School of Education, Hangzhou Normal University, Hangzhou, China.
E-mail addresses: r.deng@hznu.edu.cn (R. Deng), 2023111004026@stu.hznu.edu.cn (M. Jiang), 2023111004051@stu.hznu.edu.cn (X. Yu), luyuyan@stu.hznu.edu.cn (Y. Lu), liusasa@tourzj.edu.cn (S. Liu).

https://doi.org/10.1016/j.compedu.2024.105224
Received 15 August 2024; Received in revised form 9 December 2024; Accepted 11 December 2024
Available online 12 December 2024

(Korseberg & Elken, 2024; Yan et al., 2023). By substantially lowering the barrier for individuals without a technological background to utilise the generative capabilities of LLMs (Pack & Maloney, 2023), the advancements of ChatGPT and similar models have propelled generative AI (GenAI) into the forefront of educational discourse, igniting a mixed response of excitement and concern. While praised for its potential in content generation and natural language processing (Adeshola & Adepoju, 2023), concerns about academic integrity, overreliance on technology, and potential negative impacts on essential skills such as writing, coding, and problem-solving have fuelled apprehension among educators (Tlili et al., 2023; Wise et al., 2024) and the public (L. Li, Ma, et al., 2024; Na et al., 2024).

Despite these concerns, ChatGPT's potential to transform learning experiences and reinforce educational outcomes cannot be disregarded. Teaching and learning are prime areas for disruption by this technology (Chiarello et al., 2024; Dwivedi et al., 2023; Lian et al., 2024). Unlike traditional chatbots, which often rely on predefined responses and limited interaction patterns, ChatGPT's generative capabilities allow for more dynamic, context-aware conversations that can adapt to diverse educational scenarios (Hyun Baek & Kim, 2023; Niloy et al., 2024; Yang & Li, 2024). This capability to generate nuanced and personalised feedback represents a significant departure from earlier AI technologies that primarily focus on simple task completion (Su et al., 2023). This review focuses exclusively on ChatGPT and its impact on student learning due to its widespread recognition and familiarity among students (Hamerman et al., 2024). Furthermore, ChatGPT's output is coherent and comparable in quality to that produced by humans (Flodén, 2024; Gencer & Gencer, 2024; Jarry Trujillo et al., 2024; Lin & Chen, 2024), and its performance in delivering depth and accurate answers can surpass that of other GenAI tools (Dihan et al., 2024; Sallam et al., 2024; Williams, 2024).

The incorporation of ChatGPT into educational settings is occurring at an unprecedented pace, with a growing trend of student usage of ChatGPT for various academic work (Jo, 2023; Playfoot et al., 2024). Additionally, instructors actively seek to leverage ChatGPT in teaching and assessment practices, aiming to improve work efficiency (Laak & Aru, 2024; Shin & Lee, 2024) and optimise student performance (Bower et al., 2024). Despite calls for empirical research into the effectiveness of ChatGPT in enhancing student learning (e.g. Law, 2024; Lee & Song, 2024), a significant knowledge gap exists regarding its impact across different learning outcomes. The failure to understand these impacts before widespread implementation could diminish the quality of learning, perpetuate educational inequalities, and raise concerns about academic integrity (Jensen et al., 2024; Jiang et al., 2024). This study addresses this gap by aggregating the relevant experimental evidence to gain a holistic understanding of the field and identify research trends and promising directions for future research.

This systematic review and meta-analysis seeks to answer the fundamental question: What insights does the experimental evidence provide regarding the impact of ChatGPT on various dimensions of student learning? Gaining a nuanced understanding of this impact is crucial for informing teaching strategies and shaping the responsible implementation of ChatGPT and similar GenAI tools in educational contexts. If the evidence substantiates the positive influence of ChatGPT, it becomes essential to incorporate innovative pedagogical practices that adequately prepare students for a GenAI-driven workforce (Chan, 2023; Dianova & Schultz, 2023), such as fostering their prompt engineering skills (Walter, 2024). Positive evidence would also indicate the need for changes in teacher preparation, as instructors will need to adapt to new technologies rather than risk being left behind (Garofalo & Farenga, 2024). Conversely, negative evidence would necessitate a more cautious approach, highlighting potential areas of concern and the need for further research to mitigate any adverse effects. This review aims to provide insights for researchers, instructors, and policymakers navigating the integration of this transformative technology into the educational landscape.

2. Literature review

2.1. Student and instructor perceptions and attitudes towards ChatGPT in education

The perceptions and attitudes of students and instructors play a crucial role in shaping the adoption and integration of ChatGPT. Empirical investigations have been conducted to understand students' (Niloy et al., 2024) and instructors' (Al-khresheh, 2024) acceptance of integrating ChatGPT into teaching and learning. Generally, undergraduate (Tu & Hwang, 2023) and postgraduate students (Dai et al., 2023) demonstrate a favourable reception towards ChatGPT, despite the absence of adequate guidelines for its use (Adams et al., 2023; Zou & Huang, 2023). While major student populations embrace ChatGPT, a discernible cohort remains indifferent or cautious about its integration into education (Šedlbauer et al., 2024). A prevailing view among students is that ChatGPT can be leveraged for ideation and cognitive offloading of routine tasks rather than as a tool for automated writing completion (Barrett & Pack, 2023; Yan, 2023). Notwithstanding the occurrence of illogical, problematic, and contradictory responses (Stojanov, 2023; Urhan et al., 2024), students often exhibit trust in the capacity of ChatGPT to deliver accurate answers (Ding et al., 2023).

Instructors have exhibited ambivalent attitudes towards the use of ChatGPT in educational settings, reflecting both concerns and opportunities for pedagogical enhancement. A subset of instructors have voiced concerns regarding the potential adverse impact on education posed by the promotion of ChatGPT for students (Nam & Bai, 2023), particularly the potential dissemination of misinformation (Su & Yang, 2023), the lack of evidence supporting the obtained results (Cooper, 2023), the fostering of copying and pasting habits (Garcia Castro et al., 2024), and the impediment to developing higher-order thinking skills (Mohamed, 2024). Concurrently, instructors have posited that responsible use of ChatGPT holds promise in augmenting teaching and learning (Yusuf et al., 2024). The perception of this group is that ChatGPT exhibits efficiency in facilitating output generation across diverse academic tasks, including the development and optimisation of course plans (Okulu & Muslu, 2024), rubrics (Cooper, 2023), quiz questions (U. Lee, Chen, et al., 2024), presentation slides (Galindo-Domínguez et al., 2023), content knowledge (Su & Yang, 2023), and innovative pedagogies (Yeh, 2024). They anticipate that ChatGPT has the potential to transform students from passive knowledge recipients into active investigators (Jeon & Lee, 2023) but often remain critical regarding its implementation in areas such as assessment and feedback (ElSayary, 2023).

ChatGPT advancements have sparked enthusiasm and concern, highlighting the necessity for empirical inquiries into its potential impact. Compared to instructors, students tend to be more enthusiastic about incorporating ChatGPT into education (Chan & Tsi, 2024). While student and instructor perceptions and attitudes provide valuable insights into initial perspectives on ChatGPT in education (Chan & Lee, 2023; Monib et al., 2024; Moorhouse & Kohnke, 2024), they do not offer empirical evidence of its *actual* impact on learning outcomes. To move beyond subjective opinions, it is imperative to turn to empirical research that can establish the relationship between ChatGPT adoption and student learning. Cross-sectional and experimental studies are commonly used approaches to identify this relationship.

2.2. Strengths and limitations of cross-sectional studies in understanding ChatGPT's impact

Cross-sectional studies provide valuable insights into the early exploration of ChatGPT in education. The frequent occurring topics are students' perceptions and experiences of ChatGPT usage (Gao et al., 2024; Grájeda et al., 2024), individual differences in ChatGPT perceptions and experiences based on personal characteristics (e.g. gender; Almazrou et al., 2024; Ofem et al., 2024), and factors driving students' intentions (Jo, 2024; Maheshwari, 2024; Tan et al., 2024) and actual use of ChatGPT (Grassini et al., 2024; Salifu et al., 2024; Wijaya et al., 2024). By capturing a snapshot of student perceptions, behaviour, and performance at a specific point in time, cross-sectional studies are advantageous due to their ability to quickly gather large-scale data from diverse student populations. They are an effective starting point for identifying patterns in how learners perceive and interact with ChatGPT, as well as exploring associations between ChatGPT use and learning-related factors, such as academic performance.

While cross-sectional studies can be used to identify the relationship between the use of ChatGPT and learning outcomes, they often fail to clarify the directionality of this relationship. For example, Shahzad et al. (2024) conduct a cross-sectional study involving 362 university students in China, employing structural equation modelling to explore the relationship between ChatGPT use and academic performance. The study identifies a positive association between ChatGPT use and improved performance, thereby recommending that higher education institutions should consider utilising GenAI tools to enhance student learning. However, the cross-sectional nature of the research makes it difficult to determine whether increased ChatGPT use leads to better academic performance or if students with stronger academic results are more inclined to use ChatGPT. This ambiguity regarding the directionality of the relationship suggests that further experimental studies are necessary to establish causality. Despite the directionality problem, it is not uncommon for researchers to claim a positive effect of ChatGPT on learning performance based solely on cross-sectional evidence (e.g. Al-Qaysi et al., 2024; Boubker, 2024; Dahri et al., 2024).

Another example of the limitations of cross-sectional studies can be observed in a study by Crawford et al. (2024), which analyses data from 387 undergraduate and postgraduate students. The study employs structural equation modelling to analyse the relationships between various factors, including the use of ChatGPT and self-reported academic performance. Crawford et al. (2024) observe a negative yet statistically insignificant correlation between ChatGPT use and performance, thereby cautioning against the possible adverse impact of ChatGPT. However, apart from the potential effect of social desirability bias (Paulhus, 1991) and recall bias (Coughlin, 1990) on the accuracy of self-reported performance, cross-sectional research remains uncertain as to whether the use of ChatGPT contributes to poor academic performance or vice versa (e.g. Al-Mamary et al., 2024). Gaining insights into the directionality of this relationship is pivotal for informing educational policies and practices. If ChatGPT leads to decreased learning outcomes, interventions may be necessary to promote responsible usage and mitigate potential adverse effects. However, if students with lower performance are more inclined to use ChatGPT, the focus might shift towards providing additional support to improve their success.

The directionality problem is not the only limitation faced by cross-sectional studies examining ChatGPT's impact. These studies may be susceptible to spurious correlation (Haig, 2003), where the observed negative correlation between the use of ChatGPT and lower performance could be influenced by a confounding factor like prior knowledge. For example, individuals with limited prior knowledge may encounter difficulties comprehending the learning material and resort to ChatGPT for assistance, resulting in lower academic performance. Moreover, there might be inherent self-selection bias (Titus, 2006), as students who opt to use ChatGPT could differ systematically from those who do not. They might be more prone to overreliance on GenAI tools and a decrease in effort, potentially leading to biased results and inaccurate conclusions about the impact of ChatGPT. Lastly, cross-sectional studies often fail to contextualise ChatGPT usage within a unique learning environment, as data are typically collected from students across various educational settings through non-probabilistic convenience sampling (e.g. Acosta-Enriquez et al., 2024; Bouteraa et al., 2024; Budhathoki et al., 2024). Contextual factors, such as the curriculum structure, assignment types, and the integration of supplementary instructional technologies (Biggs & Tang, 2011), potentially influence student learning with ChatGPT and are, therefore, crucial for interpreting how and why ChatGPT impacts learning.

Although cross-sectional studies are often economical and time-saving, and contribute to identifying potential associations between ChatGPT adoption and learning outcomes (e.g. Jaboob et al., 2024; Ngo et al., 2024), they are limited in their ability to establish causal relationships or determine the temporal sequence of events. Additionally, they are prone to spurious correlations and may lack the context of specific learning environments. The limitations inherent in the cross-sectional research design emphasise the necessity for experimental studies to determine the impact of ChatGPT adoption on student learning.

2.3. Experimental studies related to ChatGPT's impact on student learning

Cross-sectional studies, while useful for identifying correlations between variables such as ChatGPT use and learning outcomes, are limited in establishing causal relationships because they assess data at a single point in time, lack the ability to determine the temporal sequence of events, and are susceptible to spurious correlation. In contrast, experimental studies offer stronger causal evidence by

Table 1
Previous reviews on ChatGPT in education.

Publication date	Author(s)	Database searched	End of the data collection period	Number of articles reviewed	Key contributions
April 2023	Lo (2023)	Academic Search Ultimate, ACM Digital Library, Education Research Complete, ERIC, IEEE Xplore, Scopus, Web of Science, and Google Scholar	February 2023	50	ChatGPT's applications and limitations in education
June 2023	Perera and Lankathilaka (2023)	ScienceDirect, Springer, Web of Science, Taylor & Francis, ResearchGate, EBSCOhost, and major academic publishers	May 2023	8	Use of ChatGPT and potential impact on higher education
July 2023	Grassini (2023)	Google Scholar and Scopus	May 2023	*Not reported*	Benefits and challenges of ChatGPT integration in education
July 2023	İpek et al. (2023)	ScienceDirect, ERIC, Wiley, Springer, Sage, Taylor & Francis, MDPI, and JSTOR	February 2023	40	Implications and concerns regarding the use of ChatGPT in education
July 2023	Jahic et al. (2023)	Google Scholar, IEEE Xplore, ScienceDirect, and Web of Science	*Not reported*	41	ChatGPT's educational applications, advantages, and disadvantages
July 2023	Montenegro-Rueda et al. (2023)	Web of Science, Scopus, and Google Scholar	June 2023	12	ChatGPT's impact on education
July 2023	Vargas-Murillo et al. (2023)	Scopus, ScienceDirect, ProQuest, IEEE Xplore, and ACM Digital Library	*Not reported*	16	ChatGPT's applications, challenges, opportunities, and impact in education
August 2023	Imran and Almusharraf (2023)	Scopus, ScienceDirect, PubMed, and Web of Science	May 2023	30	Opportunities and challenges of using ChatGPT for academic writing
August 2023	Pradana et al. (2023)	Google Scholar	*Not reported*	93	Key contributors, subtopics, and emerging research directions in ChatGPT's educational applications
September 2023	Dempere et al. (2023)	PubMed, Web of Science, IEEE Xplore, Scopus, Google Scholar, ACM Digital Library, ScienceDirect, JSTOR, ProQuest, SpringerLink, EBSCOhost, and ERIC	*Not reported*	143	ChatGPT's potential and limitation in higher education
October 2023	Ansari et al. (2024)	Google Scholar, Taylor & Francis, Emerald, Sage, Elsevier, ScienceDirect, and PubMed	May 2023	69	Use of ChatGPT in higher education
December 2023	Polat et al. (2024)	Scopus	July 2023	212	Trends, themes, and contributors of ChatGPT research in education
December 2023	Zhang and Tur (2023)	Web of Science, Scopus, ERIC, SpringerLink, IEEE Xplore, and ACM Digital Library	*Not reported*	13	ChatGPT's potential and limitation in K-12 education
February 2024	Mai et al. (2024)	Scopus, ERIC, and Google Scholar	December 2023	51	Strengths, weaknesses, opportunities, and threats of the use of ChatGPT in education
February 2024	Mahrishi et al. (2024)	Scopus	December 2023	109	ChatGPT's development and emerging dynamics in research and education
February 2024	Wong et al. (2024)	Web of Science and Altmetric	August 2023	175	Correlation between media attention and scholarly citations in ChatGPT-related education research
April 2024	Bhullar et al. (2024)	Scopus	May 2023	47	ChatGPT's applications, challenges, opportunities, and impact in higher education
May 2024	Yun and Surianshah (2024)	Scopus	*Not reported*	58	Trends, contributors, and themes of ChatGPT research in education
June 2024	Ali et al. (2024)	Academic Search Premier, Web of Science, and IEEE	October 2023	112	Benefits and limitations of ChatGPT in teaching and learning
July 2024	Amarathunga (2024)	Scopus	May 2024	45	Trends, contributors, and gaps of ChatGPT research in education
July 2024	Baig and Yadegaridehkordi (2024)	Emerald, ERIC, MDPI, SAGE, Elsevier, SpringerLink, Frontiers, PLoS ONE, Wiley, and Taylor & Francis	January 2024	57	Trends, measures, applications, and limitations of ChatGPT research in higher education
August 2024	Samala et al. (2024)	Web of Science and Scopus	December 2023	453	ChatGPT's applications, advantages, limitations, ethics considerations, and prospects

Note: If publication date is not provided, the acceptable date is recorded.

directly measuring the effects of interventions in controlled settings (Gorard & Cook, 2007). Since the introduction of ChatGPT, discrete experimental studies have been conducted to understand the impact of the technology on learning. These studies span a wide range of subjects, such as language (Maghamil & Sieras, 2024), programming (Donald et al., 2024), and health (Svendsen et al., 2024). They also cover various educational levels, such as primary schools (Almohesh, 2024), high schools (Kim, 2024), and universities (Xue et al., 2024). This diversity provides a broad perspective on ChatGPT's potential impact across diverse learning environments. However, contradictory findings have emerged, with some studies reporting significant improvements in learning outcomes (Emran et al., 2024; Lyu et al., 2024), while others indicate significant negative impacts (Shin et al., 2024; Zhang et al., 2024) or no measurable effects (Bašić et al., 2023; Farah et al., 2023). Considering the mixed findings, there remains a need for comprehensive review and meta-analysis to resolve existing contradictions and provide clearer guidance for educational practice. The following sections explore existing reviews of ChatGPT in education.

2.4. Existing reviews of ChatGPT in education

Table 1 presents a chronological list of review articles on ChatGPT in education. Most reviews provide comprehensive analyses of the technology's implications for education (Chen et al., 2024; Grassini, 2023; Lo, 2023), whereas a few adopt a more focused approach by examining specific educational stages, such as higher education (Ansari et al., 2024; Baig & Yadegaridehkordi, 2024; Perera & Lankathilaka, 2023), or addressing particular applications of ChatGPT, like as a writing assistant (Imran & Almusharraf, 2023). These existing reviews provide valuable insights into the emerging trends, potential benefits, and challenges associated with integrating ChatGPT in educational contexts. They recognise the transformative potential of ChatGPT for personalising learning experiences, enhancing student engagement, and supporting diverse learning needs. Simultaneously, the reviews consistently raise concerns regarding academic integrity, accuracy, and bias, highlighting the need for careful consideration and responsible implementation strategies to mitigate potential risks.

While the existing reviews contribute to a broader understanding of ChatGPT in education, they lack a comprehensive analysis of experimental evidence. These reviews predominantly rely on theoretical discussions, opinions, and limited cross-sectional studies that do not establish causal relationships between the adoption of ChatGPT and student learning. For instance, while some reviews highlight concerns that overreliance on ChatGPT may impede critical thinking skills (Perera & Lankathilaka, 2023; Samala et al., 2024; İpek et al., 2023), others posit that it can foster such skills by providing a platform for exploring ideas and engaging in deeper thinking and analysis (Jahic et al., 2023; Montenegro-Rueda et al., 2023). This calls for a systematic review and meta-analysis of experimental studies to evaluate the impact of ChatGPT on student learning and provide educators with evidence-based guidance. To provide conceptual clarity, this review operationally defines *student learning* as the measurable improvement in cognitive, emotional, and psychological outcomes that result from ChatGPT interventions, as assessed through various methods such as standardised tests, performance tasks, and self-evaluations. The review defines *the impact of ChatGPT* as including not only the effects of the standard, ready-to-use ChatGPT application, but also those of educational tools that use ChatGPT via APIs and other custom methods on student learning.

2.5. Research opportunities

The extant literature on ChatGPT in education often exhibits homogeneity in its scope, revealing an emerging interest in its applications while also recognising both the benefits and challenges. To move beyond conceptual exploration (e.g. García et al., 2024; Lambert & Stevens) and understand the *actual* impact of ChatGPT on student learning, examining experimental research is essential for establishing causality and drawing robust conclusions about its effectiveness in education (Ansari et al., 2024). This review aims to aggregate findings from experimental studies on ChatGPT's impact on student learning, providing a deeper understanding of its integration into educational settings and supporting evidence-based decision-making regarding its adoption. Teaching and learning do not occur in decontextualised settings (Deng et al., 2019). Understanding the context of interventions is essential for interpreting trends and patterns in experimental studies examining ChatGPT's impact and their broader implications (McGrath et al., 2024). Accordingly, this review investigates both the *characteristics* and *impacts* of ChatGPT interventions and proposes two research questions.

Research question 1 (RQ1). What are the educational stages, subject areas, intervention settings, durations, and application modes of ChatGPT interventions in experimental studies?

Research question 2 (RQ2). What are the differential effects of ChatGPT interventions on various dimensions of student learning?

Given the importance of methodological rigour for drawing valid conclusions about the impact of ChatGPT on student learning (Lo et al., 2024; Wong et al., 2024), this review examines the *methodological quality* of studies alongside intervention characteristics and reported impacts. Since this is not a methodological review, comprehensively analysing and discussing all aspects of methodological indicators for each reviewed publication is not feasible; however, certain methodological details are crucial and directly impact the validity of the experimental evidence. The review specifically focuses on sample size estimation and baseline difference control, which are considered essential for evaluating the quality of experimental studies (Shadish et al., 2002) and commonly featured in quality assessment guidelines for experimental research (Kmet et al., 2004; NHLBI, 2021). Adequate sample size ensures sufficient statistical power to detect meaningful effects, thus preventing the oversight of true effects due to insufficient power (Abraham & Russell, 2008; Peng et al., 2012). Appropriate baseline difference control increases precision in estimating experimental effects, ensuring that

observed effects can be more accurately attributed to the intervention rather than pre-existing differences between groups (Critical Appraisal Skills Programme, 2023; Howitt & Cramer, 2017). This consideration is particularly vital regarding emerging technologies such as ChatGPT, where there is a high potential for both substantial educational advantages and unintended repercussions. Therefore, the review proposes a third research question.

Research question 3 (RQ3). How do experimental studies of ChatGPT interventions determine sample size and control for baseline differences?

3. Methodology

3.1. Information source and search strategy

This systematic review and meta-analysis followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (Page et al., 2021). The review process involved developing a comprehensive search strategy, defining clear inclusion and exclusion criteria, and rigorously identifying relevant publications.

Two authors were involved in the article search process. Initial explorations across existing reviews and discrete empirical studies revealed diverse terminology used in research on ChatGPT's impact on student learning. The first and fifth authors developed a comprehensive search string, as presented in Table 2, to capture the variability by conducting preliminary searches and extracting keywords from articles published in reputable education journals, such as *Computers & Education*. Synonyms and related terms were also included to broaden the scope and identify a larger number of relevant studies. The search string employed a Boolean logic approach, combining terms related to ChatGPT (e.g. 'GPT-3.5'), experimental research methods (e.g. 'randomised controlled trial*'), and the educational context (e.g. 'student*'). This search string was used for systematic searches in EBSCOhost, IEEE Xplore, PsycINFO, Scopus, and Web of Science, and with the data retrieved up to August 31, 2024.

Guided by the research questions, specific inclusion criteria were established to ensure the relevance of selected studies (Table 3). Articles had to meet all predetermined criteria to be included. These criteria focused on studies that directly address the impact of ChatGPT on student learning, use experimental or quasi-experimental designs, and are published in peer-reviewed journals or conference proceedings in English since December 2022, shortly following ChatGPT's release to the market in November 2022. This ensured the selection of high-quality, recent research aligned with the scope of this review. Table 4 displays the exclusion criteria. Any studies that meet these criteria were excluded. Studies were excluded if they employed qualitative methodologies, examined assistive AI or non-generative AI applications, focused on outcome variables beyond student learning, or conducted comparative analyses without measuring learning outcomes. Studies that included an experimental group in which participants used additional GenAI tools, as well as those adopting a within-subject design, were also excluded. This selection process ensured the review centred on the core research question.

The systematic review and meta-analysis process adhered to the PRISMA guidelines, as illustrated in Fig. 1. An initial search yielded 1,683 records from multiple databases: Web of Science (585), EBSCOhost (226), Scopus (637), PsycINFO (30), and IEEE Xplore (205). After removing 656 duplicate records, 1027 records remained for screening. At this stage, 899 records were excluded based on title and abstract, leaving 128 reports sought for retrieval. Of these, five reports were unretrievable. Upon full-text screening of the remaining 123 reports, 51 were excluded. After a thorough quality appraisal, an additional three records were removed, resulting in 69 studies included in the final review. Of these 69 articles, 62 were meta-analysed to investigate the impact of ChatGPT interventions on five most frequently occurring outcome variables. The rigorous search and selection process ensured that the review included relevant and high-quality studies.

3.2. Quality appraisal

The quality of the articles was evaluated based on the Standard Quality Assessment Criteria (SQAC) developed by Kmet et al. (2004). SQAC is a versatile tool and can be used to evaluate the quality of both quantitative and qualitative studies. It has been widely used in existing educational review research and has demonstrated its effectiveness in assessing the quality of empirical studies (e.g. Hehir et al., 2021; Schott et al., 2020). The quantitative checklist and scoring manual were employed. Three authors participated in the quality appraisal of the publications. To ensure objectivity and reliability, the second and third authors independently assessed the quality of the publications. An interrater agreement coefficient of 0.88 demonstrated a high level of consistency in the assessment process (McHugh, 2012). Discrepancies between the two authors' assessments were resolved through discussion and by consulting the

Table 2
Search string keywords.

Terms related to ChatGPT		Terms related to experimental research methods		Terms related to the educational context
'ChatGPT' OR 'chat generative pre-trained transformer' OR 'GPT-3.5' OR 'GPT-4' OR 'GPT-4o' OR 'generative AI' OR 'generative artificial intelligence' OR 'GenAI' OR 'generative model' OR 'artificial intelligence generated content' OR 'AIGC' OR 'AI-generated'	AND	'experiment*' OR 'randomised controlled trial*' OR 'randomised controlled trial*' OR 'RCT*' OR 'quasi-experiment*' OR 'intervention'	AND	'education*' OR 'student*' OR 'learner*'

Table 3
Inclusion criteria.

Inclusion criteria	Rationale
The study used ChatGPT in the intervention.	The criterion ensures that only studies focusing specifically on ChatGPT (as opposed to other GenAI tools) are included, allowing for a targeted analysis of its impact on student learning. Given ChatGPT's distinct generative capabilities, it is crucial to isolate its effects from those of other GenAI tools (Jošt et al., 2024).
The study used experimental and quasi-experimental designs.	The criterion prioritises studies employing experimental and quasi-experimental designs, specifically those using treatment and control group comparisons, to enable the drawing of causal inferences regarding the impact of ChatGPT on student learning.
The study used students as participants.	This criterion ensures that the investigation focuses on the impact of ChatGPT on student learning, rather than on professionals (e.g. Noy & Zhang, 2023).
The study included at least one control group that did not use ChatGPT (or ChatGPT-supported learning applications) and one experimental group that did.	The inclusion of a control group ensures that any detected effect can be attributed to the impact of ChatGPT interventions rather than alternative explanations, such as developmental gains and testing effects.
The study investigated the impact of ChatGPT on cognitive (e.g. knowledge acquisition), emotional (e.g. enjoyment), and psychological outcomes (e.g. self-efficacy).	The criterion provides a holistic understanding of how this technology influences various aspects of student learning. Focusing on cognitive, emotional, and psychological dimensions ensures that the review captures the multifaceted effects of ChatGPT in education.
The study explored the differential impact of ChatGPT across various age groups and educational levels.	The criterion allows for the assessment of ChatGPT's effects across diverse learner populations, recognising that its impact may vary by age or educational context.
The study is restricted to peer-reviewed journal articles and conference papers.	The criterion prioritises the quality and rigour of research, ensuring the reliability and validity of findings. Peer-reviewed sources ensure a baseline level of methodological quality and rigour.
The study has a publication date of December 2022 or later.	The criterion acknowledges the significant influence of ChatGPT's release on the advancement and popularisation of GenAI. Studies before this date would not capture the specific dynamics of this version of GenAI.
The study is published in English.	The criterion ensures consistency and feasibility within the scope of this review. While this may limit the generalisability to non-English contexts, it aligns with the researchers' expertise.

Table 4
Exclusion criteria.

Exclusion criteria	Rationale
The study exclusively focused on the analysis of rich qualitative data (e.g. Liu et al., 2024), departing from the quantitative approach reliant on statistical analysis and hypothesis testing.	The review prioritises quantitative studies with measurable outcomes, as these allow for statistical analysis and hypothesis testing to assess ChatGPT's impact on learning. Qualitative data, while valuable, falls outside the scope of this review.
This study investigated the broader impact of assistive and non-generative AI (e.g. Tai, 2024), moving beyond the specific case of ChatGPT.	The impact of assistive and non-generative AI technologies, such as voice assistants designed for understanding and responding to requests, falls outside the scope of this review. These technologies do not possess ChatGPT's distinctive generative capabilities. However, studies investigating the effects of conversational chatbots explicitly integrating ChatGPT were included (e.g. Ng et al., 2024).
The study investigated outcome variables beyond student learning (e.g. Cingillioglu et al., 2024).	This review focuses exclusively on the direct impact of ChatGPT on learning outcomes. Studies assessing factors unrelated to student learning, such as enrolment decisions, are not directly relevant to our primary objective of evaluating the effects of ChatGPT on learning and are therefore excluded from this review.
The study compared ChatGPT to human instructors (e.g. Steiss et al., 2024) or other AI tools (e.g. Seth et al., 2024), without evaluating subsequent impacts on learning.	Mere comparisons between ChatGPT and human instructors or other technologies without direct evaluation of learning outcomes do not contribute to understanding ChatGPT's educational impact.
The study included an experimental group in which participants not only used ChatGPT but also additional GenAI tools (e.g. Saritepeci & Yildiz Durak, 2024).	The review separates the effect of ChatGPT on student learning from other GenAI tools.
The study adopted a within-subject design (e.g. Çelik, Yangın Ersanlı, & Arslanbay, 2024).	Studies adopting a within-subject design introduce potential issues such as order effects (e.g. students may perform better in later tasks regardless of ChatGPT's influence due to increased comfort with the task setup) and carryover effects (e.g. using ChatGPT after a non-ChatGPT condition may enhance subsequent performance on similar tasks because of the cognitive benefits and skills acquired during the initial non-ChatGPT task).

first author.

Each publication was scored on 14 dimensions: (1) research question or objective, (2) study design, (3) participant selection, (4) description of participant characteristics, (5) random assignment, (6) investigator blinding, (7) participant blinding, (8) outcome measures, (9) sample size, (10) analytic method, (11) estimate of variance, (12) controlling of confounders, (13) reporting of results,

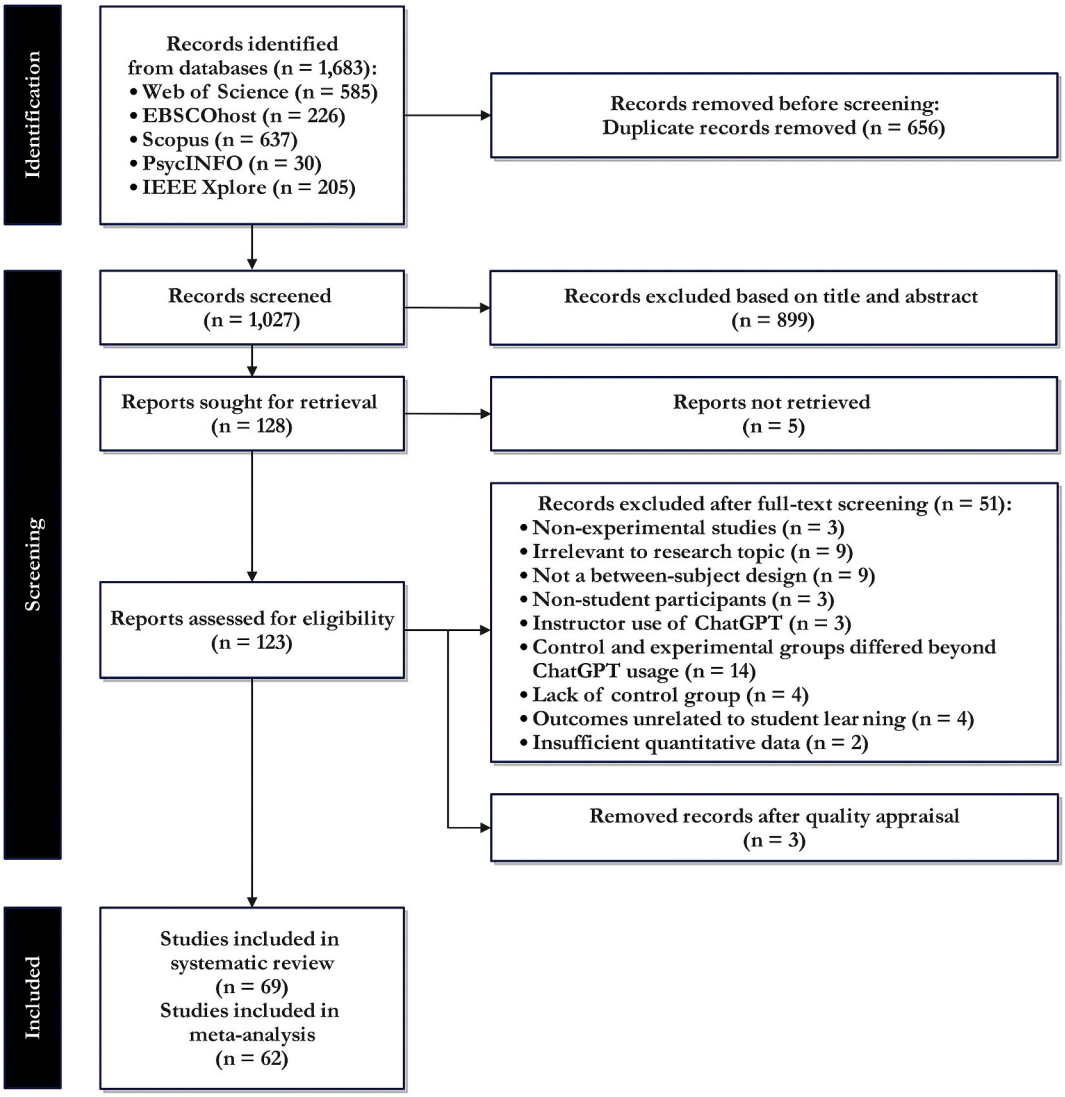


Fig. 1. Article search and selection process.

and (14) conclusions. Each dimension was scored on a scale of Yes = 2, Partial = 1, and No = 0. The option 'Not applicable' was not used in this study, as all dimensions were relevant to the reviewed publications. To ensure that the raters reached a consensus on the scoring process, the first author developed a table containing examples based on the assessment criteria (Kmet et al., 2004) to explain how each dimension should be rated. For instance, the first dimension pertains to whether the research question or objective is described. The two raters were instructed to select 'yes' when specifics about the research objectives, subjects, and interventions were reported in the introduction or the first paragraph of the methods, 'partial' when these specifics were reported in other parts of the article, and 'no' when they were not reported or were incomprehensible. To facilitate understanding, examples were provided to supplement the explanation.

The quality score for each publication was calculated by dividing the total score by the maximum possible score of 28. Higher scores indicate higher quality. Based on previous research findings, final mean scores lower than 0.50 indicates inadequate quality, while scores ranging from 0.50 to 0.70 indicate adequate quality, scores between 0.71 and 0.80 indicate good quality, and scores higher than 0.80 indicate strong quality (Lee et al., 2008). Considering the nascent research stage, the inclusion threshold was set at a conservative cut-off point of 0.5. The quality scores of the records included in this review ranged from 0.50 to 0.89, with an average score of 0.68. Three publications were excluded due to quality reasons, resulting in a total of 69 articles for the systematic review.

3.3. Data extraction

This review followed the recommendations of Higgins et al. (2019) for data extraction. Two authors were involved in the data extraction process. The first author thoroughly read 75% of the publications, extracting key information relevant to the research questions and pertinent details (e.g. research objectives, statistical analysis methods), and conducting initial coding. To facilitate collaboration, an online table was created on a knowledge management and collaboration platform, where the publication PDFs were uploaded. The first author then highlighted relevant excerpts in the publication PDFs and linked these highlights to the table where the information was documented. For example, 'We invited the second-year master's students from the University Department of Forensic Sciences, to voluntarily participate in research on essay writing as a part of the course … ' (Bašić et al., 2023, p. 2), the author highlighted this sentence in the publication PDF, recorded the code 'University' in the 'Educational stage' column of the table, and included a link to the highlighted section as a URL.

This approach established a clear connection between the codes in the table and their corresponding sources, facilitating data extraction and verification. The fourth author read the same 75% of the publications to validate the initial coding. Discrepancies were visually emphasised using colour coding and descriptive labels. The colour coding drew attention to sections where potential disagreements were identified, while the descriptive labels provided concise justifications for these discrepancies. The two authors participated in several iterations of discussions to finalise the coding scheme, aiming to reach a consensus on the categorisation and interpretation of data. Once a consensus was reached, the colour coding and descriptive labels were removed from the table, and the consensus data were recorded. Subsequently, the fourth author read the remaining 25% of the publications and extracted key information. The first author then reviewed and verified the extracted information. This collaborative approach ensured consistency in the data extraction and coding process. The coding results were presented in a tabular format.

The key variables extracted for RQ1 are educational stage, subject area, intervention setting, duration, and application mode. Educational stage denotes the level of education at which the ChatGPT intervention is implemented, classified as either university or K-12. Subject area refers to the field of study in which the intervention takes place (UNESCO Institute for Statistics, 2012). The review identified specific subject areas where ChatGPT interventions took place and grouped them into broader categories, such as arts and humanities, health and medical sciences, science, and social sciences. Intervention setting indicates the environment where the ChatGPT intervention is conducted, categorised into classroom or laboratory. Duration describes the length of time over which the ChatGPT intervention is applied, measured in units such as days, weeks, or months. The review categorised duration into specific timeframes, including less than 1 week, 1–4 weeks, 5–10 weeks, and more than 10 weeks. Application mode is defined as the way ChatGPT is utilised in the intervention, either as a direct learning tool where students interact with ChatGPT directly or as a ChatGPT-supported learning application integrated into broader educational platforms and tools to enhance the learning experience.

The key variables extracted for RQ2 are academic performance, affective-motivational states, higher-order thinking propensities, self-efficacy, and mental effort. The outcome variables were operationally defined to ensure conceptual clarity, guided by existing meta-analyses in educational settings (Brom et al., 2017; Castillo-Manzano et al., 2016; David et al., 2024; Talsma et al., 2018; Zhan et al., 2023). In this review, *academic performance* refers to students' achievements measured objectively through exams (Lyu et al., 2024), exercises (Farah et al., 2023), projects (Li, 2023), tasks (Meyer et al., 2024), and tests (Boudouaia et al., 2024), rather than through self-reported perceptions or subjective judgments of learning. *Affective-motivational states* indicate the affective and motivational conditions of learners, encompassing factors such as learning interest (Donald et al., 2024), intrinsic motivation (Chen & Chang, 2024), emotional engagement (Suciati et al., 2024), and enjoyment (Chang et al., 2024). *Higher-order thinking propensities* describe students' subjective beliefs, perceptions, or awareness related to cognitive processes that involve advanced levels of thinking, such as computational thinking (Ameen et al., 2024), creative thinking (Li, 2023), critical thinking (Darmawansah et al., 2024), problem-solving (Hu, 2024b), and reflective thinking (Essel et al., 2024). *Self-efficacy* denotes the belief in students' ability to successfully execute tasks and achieve goals in academic settings, such as programming (Donald et al., 2024; Yilmaz & Karaoglan Yilmaz, 2023). *Mental effort*, often considered a reflection of cognitive load (Paas et al., 2010), is conceptualised as the amount of cognitive resources allocated to process information and complete learning-related tasks, such as teaching aids creation (Ji et al., 2023) and reading comprehension (Wang et al., 2024).

RQ3 addresses sample size estimation and baseline difference control. Sample size estimation is the method used to statistically determine the number of participants required for the study, classified as yes or no regarding the use of power analysis to ensure sufficient sample size for detecting meaningful effects. Baseline difference control refers to the procedures implemented to account for and mitigate pre-existing differences between participants, thereby ensuring that observed outcomes can be attributed to the intervention rather than inherent group disparities. It is classified as yes or no regarding the use of measures including pre-test, covariate, and random assignment. For RQ3, the key variables extracted included the use of power analysis, pre-tests, covariates, and random assignment, each classified as either yes or no.

3.4. Data analysis

3.4.1. Overview of data analysis

To identify the overall patterns in experimental studies utilising ChatGPT interventions, this review analysed all 69 studies to address RQ1. Descriptive statistics were calculated to summarise coded elements and identify patterns across the dataset related to educational stage, subject area, intervention setting, duration, and application mode. Intervention features such as duration (Dondio et al., 2023) and participant characteristics such as educational stage (Wong & Adesope, 2021) are commonly used as moderators in meta-analyses conducted in educational settings. Consequently, these five variables not only serve as descriptive variables for RQ1 but

also as moderating variables for RQ2. To address RQ2, a series of meta-analyses were conducted to investigate the effect of ChatGPT on various dimensions of student learning. Outcome variables for RQ2 were not predetermined; rather, they emerged from an analysis of all eligible studies and were classified and identified based on their frequency of investigation. To ensure conciseness while observing patterns, the review analysed only five of the most frequently investigated learning outcome variables for RQ2; those reported sporadically in experimental studies, such as self-regulated learning (Ng et al., 2024) collaboration tendency (Darmawansah et al., 2024), and decisional conflicts (Hu, 2024a), were excluded from the meta-analysis. Of the 69 studies, 62 were meta-analysed to evaluate the extent to which ChatGPT interventions impact (1) academic performance, (2) affective-motivational states, (3) higher-order thinking propensities, (4) self-efficacy, and (5) mental effort. Similarly, these 62 studies were further analysed regarding sample size estimation and baseline difference control using descriptive statistics to address RQ3. This allowed for an interpretation of the meta-analysis results in relation to critical aspects of the methodological quality of the meta-analysed studies.

3.4.2. Meta-analysis

To address RQ2, the meta-analyses were conducted and statistically analysed using Comprehensive Meta-Analysis (CMA) 3.0 (Borenstein, 2022). Considering the inherent bias of standardised mean difference effect sizes in studies with small samples, Hedges' *g* was selected as the standard measure of effect size (Hedges, 1981). According to Hattie (2008), who analysed over 800 meta-analyses, Cohen's *d* values of 0.20, 0.40, and 0.60 are interpreted as small, medium, and large effects, respectively, in educational contexts. Hedges' *g*, closely related to Cohen's *d*, accounts for small-sample bias (Hedges & Olkin, 1985) and can be interpreted using the same effect size conventions as Cohen's *d*. Hedges' *g* was computed for each independent study using the means and standard deviations reported. When these data were not available, alternative statistical metrics, such as t-values from t-tests or F-values (Borenstein et al., 2009), were utilised to estimate the effect size. To assign larger weights to studies with greater precision or representativeness and provide an unbiased estimate of effect sizes (Borenstein et al., 2009), weighted average effect sizes (*g*+) were computed (e.g. Wong & Adesope, 2021).

A total of 18 studies examined the impact of the ChatGPT interventions on multiple outcome variables. To ensure the independence of effects, separate meta-analyses were conducted for each outcome variable. When a single study included multiple experiments, we computed an effect size for each experiment. In cases where a study reported multiple post-intervention assessments (e.g. Task 1, Task 2) without providing a summative measure or an aggregate result from the same cohort of participants, we calculated a single combined mean effect size. Treating different outcomes from the same participants as independent is not recommended, as it could lead to an erroneous estimation of the overall effect (Halme et al., 2010). Our review exhibited considerable variation in subject areas,

Table 5
Information about the source publications (n = 69).

	Frequency	Percent
Year of publication		
2024	57	82.61
2023	12	17.39
Type of publication		
Journal article	58	84.06
Conference paper	11	15.94
Source of publication		
Education and Information Technologies	6	8.70
Computers and Education: Artificial Intelligence	4	5.80
IEEE Transactions on Learning Technologies	3	4.35
Journal of Educational Computing Research	2	2.90
Journal of Computer Assisted Learning	2	2.90
International Review of Research in Open and Distributed Learning	2	2.90
International Journal of Educational Technology in Higher Education	2	2.90
International Journal of Engineering Pedagogy	2	2.90
International Conference on Innovative Technologies and Learning	2	2.90
Educational Technology & Society	2	2.90
Asia Pacific Journal of Education	2	2.90
Arab World English Journal	2	2.90
Other sources	38	55.07
Geographic region (first author)		
Asia	49	71.01
Europe	12	17.39
North America	5	7.25
South America	2	2.90
Africa	1	1.45
Study location		
Asia	39	56.52
Europe	9	13.04
North America	4	5.80
Africa	2	2.90
Not reported	15	21.74

Note: Percent may not sum to 100% due to rounding.

intervention duration, and other factors. Consequently, a random-effects model was deemed more appropriate than a fixed-effects model, as it accounts for variability in effect sizes across studies due to sampling error and differences in true effects (Borenstein et al., 2009). Subsequently, heterogeneity analyses were conducted to verify the adequacy of the chosen model.

Test for heterogeneity was assessed using the I^2 test. According to Higgins and Thompson's (2002) criteria, I^2 values of 25%, 50%, and 75% indicate low, medium, and high heterogeneity, respectively. If substantial heterogeneity was observed and the number of studies was adequate, we applied a random-effects model with moderator analyses to identify potential sources of effect differences.

Tests for publication bias We employed three methods to assess publication bias in each meta-analysis. First, funnel plots were constructed and visually inspected. To further examine the presence of funnel plot asymmetry, Begg and Mazumdar's rank correlation test (Begg & Mazumdar, 1994) was conducted. Additionally, Egger's linear regression test was carried out. When publication bias was detected, we used trim-and-fill analyses (Duval & Tweedie, 2000) to estimate the number of missing studies and adjusted the meta-analysis effects accordingly. To identify potential outliers, Grubbs' test (Grubbs, 1969) was performed.

4. Results

4.1. Overview of source publications

Basic information about the source publications is displayed in Table 5. Regarding the year of publication, most articles were published in 2024 (n = 57), with the remainder published in 2023 (n = 12), indicating a growing trend in the number of experimental studies. The majority of the publications are journal articles (n = 58), with a smaller number being conference papers (n = 11). Regarding the sources of publication, *Education and Information Technologies* (n = 6) is the most represented source, followed by *Computers and Education: Artificial Intelligence* (n = 4) and *IEEE Transactions on Learning Technologies* (n = 3). Nine journals (e.g. *Journal of Educational Computing Research*) and conferences (i.e. *International Conference on Innovative Technologies and Learning*) are each represented twice. The remaining 38 articles are sourced from a diverse range of other journals (e.g. *Computers & Education*) and conferences (e.g. *International Conference on Computers in Education*), with each appearing only once. Geographically, the first authors are predominantly from Asia (n = 49), followed by Europe (n = 12), North America (n = 5), South America (n = 2) and Africa (n = 1). The studies are conducted primarily in Asia (n = 39), followed by Europe (n = 9), North America (n = 4), and Africa (n = 2), with 15 studies not reporting their location.

The following section presents the findings for each research question. Interpretations and discussions of these findings are provided in Section 5.

4.2. RQ1: what are the educational stages, subject areas, intervention settings, durations, and application modes of ChatGPT interventions in experimental studies?

4.2.1. Educational stage

The educational stage in the reviewed publications was categorised into (1) university and (2) K-12. Of the 69 studies included in the review, 58 (84.06%; e.g. Shin et al., 2024; Silitonga et al., 2023) were conducted at the university level. Ten studies (14.49%; e.g. Khuibut et al., 2023; Ng et al., 2024) were conducted in K-12 settings. One study did not explicitly indicate whether the experiment was conducted in a K-12 or university setting (i.e. Suciati et al., 2024).

4.2.2. Subject area

The reviewed publications spanned diverse subject areas. Among the 69 studies, a substantial number focused on language education (n = 22, 31.88%; e.g. Kim, 2024; Meyer et al., 2024), making it the most frequently studied context. This is followed by computing (n = 9, 13.04%; e.g. Liao et al., 2024; Xue et al., 2024), health (n = 8, 11.59%; e.g. Gan et al., 2024; Wu et al., 2024), physics (n = 8, 11.59%; e.g. Alneyadi & Wardat, 2024; Beltozar-Clemente & Díaz-Vega, 2024), education (n = 7, 10.14%; e.g. Ji et al., 2023; Li, 2023), business (n = 3, 4.35%; e.g. Hu, 2024a; Hu, 2024b), mathematics and statistics (n = 3, 4.35%; e.g. Lu et al., 2024; Wu et al., 2023), and arts (n = 2, 2.90%; i.e. Chandrasekera et al., 2024; Zhou & Kim, 2024). Several subjects were each represented by a single study (1.45%), including agriculture (Donald et al., 2024), engineering (Zhang et al., 2024), law (Shi et al., 2024), and life science (Bašić et al., 2023). Additionally, three studies (4.35%; e.g. Almohesh, 2024) did not specify the subject area. To facilitate the moderator analysis for RQ2, these specific subject areas were categorised into arts and humanities (n = 24, 34.78%), science (n = 21, 30.43%), social sciences (n = 11, 15.94%), health and medical sciences (n = 8, 11.59%), and others (n = 5, 7.25%).

4.2.3. Intervention setting

The intervention settings in the reviewed studies were predominantly conducted in classroom environments, with 60 out of the 69 studies (86.96%; e.g. Aydın Yıldız, 2023; Farah et al., 2023) taking place in classrooms. A smaller proportion of studies were conducted in laboratory settings (n = 6, 8.70%; e.g. Niloy et al., 2023; Stadler et al., 2024). Three studies (4.35%; e.g. Chandrasekera et al., 2024) did not report the intervention setting.

4.2.4. Duration

The intervention durations in the reviewed publications varied widely, reflecting the diverse nature of ChatGPT applications in education. Among the 69 studies, the duration ranged from a 10-min session (Zhang et al., 2024) to a 16-week semester (Gao, 2024). The units used to measure intervention duration varied across studies (e.g. days or weeks), preventing a direct comparison. For this

review, durations were reclassified into four categories: less than 1 week, 1–4 weeks, 5–10 weeks, and more than 10 weeks. Of the studies reviewed, 22 studies (31.88%; e.g. Huesca et al., 2024; Yilmaz & Karaoglan Yilmaz, 2023) fell into the category of 5–10 weeks, representing the most prevalent duration. They were closely followed by 15 studies (21.74%; e.g. Maghamil & Sieras, 2024; Urban et al., 2024) categorised under less than 1 week, 12 studies (17.39%; e.g. Kavadella et al., 2024; Ng et al., 2024) that spanned 1–4 weeks, and 10 studies (14.49%; e.g. Emran et al., 2024; Naamati-Schneider & Alt, 2024) that extended beyond 10 weeks. Notably, 10 studies (14.49%; e.g. Ironsi & Ironsi, 2024) lacked information on intervention duration.

4.2.5. Application mode

The application modes in the reviewed publications were categorised into (1) direct learning tools and (2) ChatGPT-supported learning applications. Among the 69 studies, 55 (79.71%; e.g. Ameen et al., 2024; Avello-Martínez et al., 2024; Essien et al., 2024; Kucuk, 2024) used ChatGPT as a direct learning tool. Direct learning tools refer to using the standard ChatGPT application *directly* in the learning process. For instance, students interacted with ChatGPT to generate ideas (Mahapatra, 2024), obtain explanations (Alneyadi & Wardat, 2023), clarify misconceptions (Essel et al., 2024), seek task advices (Chang et al., 2024), and receive writing feedback (Boudouaia et al., 2024). This approach leverages ChatGPT in its most direct form, providing immediate assistance and information.

Conversely, 14 studies (20.29%; e.g. H.-Y. Lee, Chen, et al., 2024; Li, 2023; Liao et al., 2024) implemented ChatGPT-supported learning applications. These involve integrating ChatGPT's functionalities into *broader* educational platforms and tools through APIs or custom implementations to offer a more interactive and seamless learning experience. Examples include gamified learning platforms that integrate ChatGPT into educational games (Chen & Chang, 2024) or programming platforms that incorporate ChatGPT to provide just-in-time coding assistance (Shang & Geng, 2024; Sun et al., 2024).

4.3. RQ2: what are the differential effects of ChatGPT interventions on various dimensions of student learning?

4.3.1. Outlier and publication bias

Grubbs' test identified one outlier concerning academic performance (Ironsi & Ironsi, 2024). However, we did not exclude this study, as no anomalies were detected regarding the intervention, measurement procedure, or calculation. In addition, we conducted an exploratory meta-analysis in the absence of this study to ensure that the overall effect was not influenced by a small number of potentially unusual studies.

Next, we examined publication bias (Table 6). We began with an exploratory analysis by visualising funnel plots (Appendix A). Certain funnel plots displayed asymmetry, which may suggest publication bias. Rank correlation tests indicated that publication bias was present in the meta-analyses for academic performance, τ (N = 51) = 0.321, $p < .001$, and affective-motivational states, τ (N = 20) = 0.484, $p = .003$, but not for higher-order thinking propensities, τ (N = 15) = 0.371, $p = .054$, self-efficacy, τ (N = 7) = 0.048, $p = .881$, or mental effort, τ (N = 4) = −0.333, $p = .497$. Additionally, we applied Egger's linear regression test, which further indicated potential publication bias in academic performance and affective-motivational states ($p > .05$). Consequently, we adjusted the meta-analytic effects for academic performance and affective-motivational states using trim-and-fill analyses.

4.3.2. Overall effects

Table 7 provides the overall analysis of the weighted means of the effect sizes for all outcome variables under a random-effects model. Forest plots of ChatGPT interventions for each outcome variable are provided in Appendix B.

Academic performance 44 out of 51 effect sizes were positive, indicating that most ChatGPT interventions enhanced academic achievement. The weighted mean effect size was $g+ = 0.712$, 95% CI [0.497, 0.926], SE = 0.109, $p < .001$, suggesting a significant large effect of ChatGPT interventions. A trim-and-fill analysis resulted in the addition of six studies, yielding an adjusted effect size of $g+_{(\text{trim-and-fill})} = 0.881$, which was slightly larger than the original effect ($g+ = 0.712$). This result suggests that the positive impact of ChatGPT interventions on academic performance might be even stronger when accounting for potential publication bias. The confidence intervals of both the original [0.497, 0.926] and adjusted estimates [0.606, 1.155] did not include zero, further supporting the consistency of this positive effect. Heterogeneity was significant, $I^2 = 91.789\%$, $Q = 608.968$, df = 50, $p < .001$, indicating considerable heterogeneity among studies, with one or more moderators potentially accounting for this mean effect.

Affective-motivational states 17 out of 20 effect sizes were positive, indicating that most studies support the idea that ChatGPT interventions promote students' affective-motivational states. The computed significant effect size was large, $g+ = 0.881$, 95% CI

Table 6
Publication bias test.

Outcomes	k	Rank correlation		Egger's linear regression test			
		Kendall's τ	p	Egger's intercept	SE	95% CI	p
Academic performance	51	0.321	<0.001	5.512	0.925	[3.653, 7.371]	<0.001
Affective - motivational states	20	0.484	0.003	4.849	1.681	[1.317, 8.381]	0.010
Higher-order thinking propensities	15	0.371	0.054	−2.021	2.514	[-7.451, 3.409]	0.436
Self-efficacy	7	0.048	0.881	1.903	5.380	[-11.928, 15.733]	0.738
Mental effort	4	−0.333	0.497	−1.397	3.215	[-15.229, 12.435]	0.706

Note: SE = standard error; 95% CI = 95% confidence interval of Egger's intercept.

Table 7
Summary of random-effects model.

Outcomes	Effect size				Heterogeneity test			
	k	$g+$	SE	95% CI for $g+$	Q	df(Q)	p	I^2 (%)
Academic performance	51	0.712***	0.109	[0.497, 0.926]	608.968	50	<0.001	91.789
Affective-motivational states	20	0.881***	0.178	[0.531, 1.231]	265.744	19	<0.001	92.850
Higher-order thinking propensities	15	0.703***	0.182	[0.345, 1.060]	144.697	14	<0.001	90.325
Self-efficacy	7	0.441	0.297	[-0.141, 1.023]	58.896	6	<0.001	89.812
Mental effort	4	−0.675*	0.304	[-1.271, −0.079]	12.179	3	0.007	75.368
Academic performance $_{\text{trim-and-fill}}$	57	0.881	–	[0.606, 1.155]	1350.694	–	–	–
Affective-motivational states $_{\text{trim-and-fill}}$	23	1.122	–	[0.685, 1.560]	628.117	–	–	–

Note: k = number of effect sizes; $g+$ = mean effect size; SE = standard error of mean correlation; Q = Cochran's homogeneity test statistic; I^2 = scale-free index of heterogeneity.
*$p < .05$, **$p < .01$, ***$p < .001$.

[0.531, 1.231], SE = 0.178, $p < .001$. A trim-and-fill analysis for affective-motivational states added three studies, yielding an adjusted effect size of $g+_{(\text{trim-and-fill})} = 1.122$, higher than the original $g+ = 0.881$. This increase indicates that the positive effects on affective-motivational states might be even greater when considering potential publication bias. Notably, the confidence intervals for both the original [0.531, 1.231] and adjusted effect sizes [0.685, 1.560] exclude zero, confirming the robustness of this positive effect. The heterogeneity tests were also significant, $I^2 = 92.850\%$, $Q = 265.744$, df = 19, $p < .001$.

Higher-order thinking propensities 14 of 15 effect sizes were positive, indicating that the majority of ChatGPT interventions significantly improved higher-order thinking propensities, such as computational thinking, critical thinking, and reflective thinking. The computed significant effect size was large, $g+ = 0.703$, 95% CI [0.345, 1.060], $SE = 0.182$, $p < .001$. The heterogeneity tests were significant as well, $I^2 = 90.325\%$, $Q = 144.697$, df = 14, $p < .001$.

Self-efficacy Five out of seven effect sizes were positive, indicating that most ChatGPT interventions increased self-efficacy. The computed significant effect size was moderate but not significant, $g+ = 0.441$, 95% CI [−0.141, 1.023], SE = 0.297, $p = .137$. The heterogeneity tests were significant, $I^2 = 89.812\%$, $Q = 58.896$, df = 6, $p < .001$.

Mental effort Three out of four effect sizes were negative, indicating that the majority of studies support the idea that ChatGPT interventions reduced mental effort. The weighted mean effect size was large, $g+ = -0.675$, 95% CI [−1.271, −0.079], SE = 0.304, p

Table 8
Moderator analyses for academic performance.

Moderator	k	$g+$	SE	95% CI for $g+$	Q_{between}	df(Q)	p_{adjusted}
Educational stage							
K-12	10	0.547***	0.118	[0.314, 0.779]			
University	41	0.754***	0.140	[0.480, 1.028]			
Between levels (Q_B)					1.283	1	0.257
Subject area							
Arts and humanities	23	1.045***	0.219	[0.615, 1.475]			
Health and medical sciences	5	0.916**	0.340	[0.250, 1.581]			
Science	14	0.354**	0.119	[0.122, 0.587]			
Social sciences	6	0.561*	0.257	[0.057, 1.065]			
Others	3	−0.012	0.445	[-0.885, 0.861]			
Between levels (Q_B)					10.478	4	0.033
Intervention setting							
Classroom	43	0.783***	0.097	[0.592, 0.974]			
Laboratory	5	−0.213	0.335	[-0.870, 0.444]			
Not reported	3	1.268	0.714	[-0.131, 2.666]			
Between levels (Q_B)					8.754	2	0.013
Duration							
<1 week	12	−0.048	0.186	[-0.413, 0.317]			
1–4 weeks	10	1.231***	0.272	[0.698, 1.765]			
5–10 weeks	15	0.913***	0.155	[0.610, 1.217]			
>10 weeks	9	0.754***	0.171	[0.420, 1.088]			
Not reported	5	0.741	0.409	[-0.061, 1.543]			
Between levels (Q_B)					21.851	4	<0.001
Application mode							
ChatGPT-supported learning application	9	0.488***	0.141	[0.212, 0.765]			
Direct learning tool	42	0.757***	0.130	[0.503, 1.012]			
Between levels (Q_B)					1.966	1	0.161

Note: k = number of effect sizes; $g+$ = mean effect size; SE = standard error of mean correlation; Q_{between} = Cochran's homogeneity test statistic for between-group heterogeneity.
$p < .05$.
$p < .01$.
$p < .001$.

= .026. The heterogeneity tests showed significant and large heterogeneity, $I^2 = 75.368\%$, $Q = 12.179$, df = 3, $p = .007$.

Overall, the results showed that ChatGPT interventions significantly enhanced academic performance ($g+ = 0.712$), affective-motivational states ($g+ = 0.881$), and higher-order thinking propensities ($g+ = 0.703$); significantly reduced mental effort ($g+ = -0.675$); but did not have a significant effect on self-efficacy ($g+ = 0.441$).

4.3.3. Moderator analyses

Separate moderator analyses were conducted for academic performance, affective-motivational states, and higher-order thinking propensities to explore possible causes of heterogeneity. Differences between the individual moderator categories were tested for significance using the 95% confidence intervals.

Moderator analysis was conducted to assess whether educational stage, subject area, intervention setting, duration, and application mode moderate the impact of ChatGPT interventions on academic performance (Table 8). The subject area significantly moderated the effect ($Q_{between} = 10.478$; df = 4; $p = .033$). Positive effects were observed for instruction in arts and humanities ($g+ = 1.045$), health and medical sciences ($g+ = 0.916$), science ($g+ = 0.354$), and social sciences ($g+ = 0.561$). Conversely, no significant effect was observed for instruction in other subject areas ($g+ = -0.012$). Another significant moderator is the intervention setting ($Q_{between} = 8.754$; df = 2; $p = .013$). ChatGPT interventions in traditional classroom settings demonstrated a significant positive impact on academic performance ($g+ = 0.783$), whereas those in laboratory settings yielded a non-significant effect ($g+ = -0.213$). Duration also emerged as a significant moderator ($Q_{between} = 21.851$; df = 4; $p < .001$). ChatGPT was found to be beneficial across various intervention durations, except for those lasting less than 1 week. Experiments with durations ranging from 1 to 4 weeks ($g+ = 1.231$) showed larger effect sizes than those lasting 5–10 weeks ($g+ = 0.913$) and more than 10 weeks ($g+ = 0.754$). No moderating effects were identified for educational stage ($Q_{between} = 1.283$; df = 1; $p = .257$) or application mode ($Q_{between} = 1.966$; df = 1; $p = .161$).

Additionally, the review carried out moderator analysis to evaluate whether educational stage, subject area, duration, and application mode moderate the impact of ChatGPT interventions on affective-motivational states (Table 9). Moderator analysis for intervention setting was not performed, as all relevant studies were carried out in classroom settings. The only significant moderator was educational stage ($Q_{between} = 11.138$; df = 2; $p = .004$). A positive effect was observed for university settings ($g+ = 1.155$), whereas no significant impact was noted for K-12 settings ($g+ = 0.378$). Additionally, no moderating effects were found for subject area ($Q_{between} = 7.275$; df = 4; $p = .122$), duration ($Q_{between} = 6.130$; df = 3; $p = .105$), or application mode ($Q_{between} = 0.272$; df = 1; $p = .602$).

The review further performed moderator analysis to examine whether subject area, intervention setting, duration, and application mode moderate the impact of ChatGPT interventions on higher-order thinking propensities (Table 10). Moderator analysis for educational stage was not performed, as all relevant studies were conducted at the university level. Duration emerged as a significant moderator ($Q_{between} = 15.471$; df = 4; $p = .004$), with studies lasting 1–4 weeks demonstrating larger effect sizes ($g+ = 1.173$) compared to those ranging from 5 to 10 weeks ($g+ = 0.499$). The effect size for studies extending beyond 10 weeks was non-significant ($g+ = 0.117$). Although ChatGPT interventions lasting less than 1 week ($g+ = 1.409$) showed a positive impact on higher-order

Table 9
Moderator analyses for affective-motivational states.

Moderator	k	g+	SE	95% CI for g+	$Q_{between}$	df(Q)	$p_{adjusted}$
Educational stage							
K-12	5	0.378	0.229	[-0.070, 0.826]			
University	14	1.155***	0.252	[0.661, 1.648]			
Not reported	1	0.028	0.233	[-0.429, 0.485]			
Between levels (Q_B)					11.138	2	0.004
Subject area							
Arts and humanities	6	0.454	0.313	[-0.161, 1.068]			
Health and medical sciences	2	1.521***	0.279	[0.975, 2.067]			
Science	5	1.076*	0.452	[0.190, 1.962]			
Social sciences	5	1.143**	0.353	[0.451, 1.835]			
Others	2	0.475	0.673	[-0.843, 1.794]			
Between levels (Q_B)					7.275	4	0.122
Duration							
<1 week	4	0.655	0.444	[-0.215, 1.526]			
1–4 weeks	9	1.432***	0.366	[0.714, 2.149]			
5–10 weeks	6	0.356	0.234	[-0.103, 0.816]			
>10 weeks	1	0.626*	0.285	[0.067, 1.186]			
Between levels (Q_B)					6.130	3	0.105
Application mode							
ChatGPT-supported learning application	8	1.007**	0.309	[0.401, 1.614]			
Direct learning tool	12	0.805***	0.233	[0.348, 1.262]			
Between levels (Q_B)					0.272	1	0.602

Note: k = number of effect sizes; g+ = mean effect size; SE = standard error of mean correlation; $Q_{between}$ = Cochran's homogeneity test statistic for between-group heterogeneity.
$p < .05$.
p < .01.
$p < .001$.

Table 10
Moderator analyses for higher-order thinking propensities.

Moderator	k	*g*+	SE	95% CI for *g*+	$Q_{between}$	df(*Q*)	$p_{adjusted}$
Subject area							
Arts and humanities	1	0.613*	0.247	[0.128, 1.097]			
Health and medical sciences	2	0.735	0.645	[-0.529, 2.000]			
Science	5	0.357*	0.155	[0.053, 0.661]			
Social sciences	7	0.955**	0.297	[0.372, 1.537]			
Others	–	–	–	–			
Between levels (Q_B)					3.490	3	0.322
Intervention setting							
Classroom	12	0.806***	0.218	[0.379, 1.233]			
Laboratory	3	0.318	0.164	[-0.003, 0.639]			
Between levels (Q_B)					3.194	1	0.074
Duration							
<1 week	1	1.409***	0.312	[0.797, 2.020]			
1–4 weeks	3	1.173*	0.577	[0.043, 2.304]			
5–10 weeks	6	0.499***	0.120	[0.263, 0.735]			
>10 weeks	1	0.117	0.163	[-0.202, 0.436]			
Not reported	4	0.658	0.416	[-0.157, 1.473]			
Between levels (Q_B)					15.471	4	0.004
Application mode							
ChatGPT-supported learning application	7	0.671**	0.242	[0.196, 1.146]			
Direct learning tool	8	0.727**	0.273	[0.192, 1.262]			
Between levels (Q_B)					0.024	1	0.878

Note: k = number of effect sizes; *g*+ = mean effect size; SE = standard error of mean correlation; $Q_{between}$ = Cochran's homogeneity test statistic for between-group heterogeneity.
$p < .05$.
$p < .01$.
$p < .001$.

thinking propensities, the limited sample size in this category (one study) did not provide enough information to determine the effect. No significant moderating effects were detected for subject area ($Q_{between} = 3.490$; df = 3; $p = .322$), intervention setting ($Q_{between} = 3.194$; df = 1; $p = .074$), or application mode ($Q_{between} = 0.024$; df = 1; $p = .878$).

4.4. RQ3: how do experimental studies of ChatGPT interventions determine sample size and control for baseline differences?

4.4.1. Sample size estimation

The sample sizes in the 62 meta-analysed studies varied considerably, ranging from as few as 18 participants (Bašić et al., 2023) to as many as 600 participants (Niloy et al., 2023), with an average sample size of 106. Sample sizes were categorised into four groups to better understand this distribution: ≤50, 51–100, 101–200, and >200. The most common sample size range was 51–100 participants (n = 24), followed by ≤ 50 (n = 19), 101–200 (n = 11), and >200 (n = 8). Power analysis, a crucial method for determining adequate sample sizes, was conducted in only five out of 62 studies (8.06%; i.e. Alneyadi & Wardat, 2023; Chen & Chang, 2024; Donald et al., 2024; Svendsen et al., 2024; Urban et al., 2024). This suggests that most studies may not have formally assessed whether their sample sizes were sufficient to detect meaningful effects.

4.4.2. Baseline difference control

The meta-analysed studies employed several key strategies to control for baseline differences, including pre-tests (n = 52), random assignment (n = 39), and covariates (n = 24).

Pre-test measures were commonly used to assess baseline equivalence between the control and experimental groups, as well as to measure baseline levels of the outcome variables before the intervention, providing essential data for statistical analyses that control for covariates (Howitt & Cramer, 2017). Among the 62 meta-analysed studies, 52 (83.87%) incorporated pre-test measures, such as conceptual tests (Chen & Chang, 2024), comprehension tests (Wang et al., 2024), and diagnostic tests (Zhou & Kim, 2024). Ten studies (16.13%; e.g. Beltozar-Clemente & Díaz-Vega, 2024; Kosar et al., 2024) did not include any pre-test measures. Of the 62 studies, 24 (38.71%; e.g. Darmawansah et al., 2024; Wu et al., 2023) employed covariates to control for potential confounding variables through ANCOVA, MANCOVA, and regression analysis, whereas the remaining 38 studies (61.29%; e.g. Chandrasekera et al., 2024; Lyu et al., 2024) did not incorporate covariates in their data analysis. Notably, 23 studies (37.10%; e.g. Hu, 2024b; Svendsen et al., 2024) included pre-test measures and covariates, and nine studies (14.52%; e.g. Ahmed Moneus & Al-Wasy, 2024; Liu et al., 2023) neither included pre-test measures nor used covariates; 29 studies (46.77%; e.g. Mugableh, 2024; Mun, 2024) included pre-test measures but did not use them as covariates, and one study (1.61%; i.e. Shang & Geng, 2024) included covariates but did not measure these covariates during the pre-intervention phase.

In addition to the use of pre-test measures and covariates, the methodological quality of these studies is further demonstrated by the implementation of random assignment. Random assignment is a key experimental technique that ensures each participant has an equal chance of being assigned to any group, thereby minimising selection bias and enhancing the study's internal validity (Bryman & Bell,

2018). Among the 62 studies, 39 (62.90%; e.g. Y. Li, Ma, et al., 2024; Shi et al., 2024) employed random assignment to allocate participants to different groups. Of these 39 studies, 16 (41.03%; e.g. Wahba et al., 2024; Xiao, 2024) included pre-test measures and covariates, strengthening the reliability of their findings by controlling for baseline differences and potential confounders. Conversely, three studies (4.84%; e.g. Ahmed Moneus & Al-Wasy, 2024) did not include pre-test measures, covariates, or random assignment.

5. Discussion

This section interprets the results in relation to the research questions, in conjunction with existing empirical observations and review studies. The key findings for each research question are summarised in Fig. 2 and discussed in greater detail below.

5.1. RQ1: what are the educational stages, subject areas, intervention settings, durations, and application modes of ChatGPT interventions in experimental studies?

Exploring RQ1 indicated that ChatGPT interventions are predominantly implemented in university-level classrooms, focusing on language education. These interventions typically last several weeks and use ChatGPT as a direct learning tool. These key findings are discussed in more detail.

First, among the 69 reviewed studies, 84.06% are conducted at the university level, whereas 14.49% focus on K-12 settings. This observation provides direct evidence that ChatGPT interventions are predominantly conducted among university students. A mix of excitement and trepidation among academics (Islam & Islam, 2024), teaching staff (Jochim & Lenz-Kesekamp, 2024), and administrators (Korseberg & Elken, 2024) in higher education institutions may have contributed to this trend. It is also possible that this trend is influenced by the limited accessibility of computers and mobile devices for K-12 students (Gao et al., 2014), as their usage is generally less prevalent than in higher education. The complexity of learning objectives at the university level not only requires students to master foundational knowledge and skills but also involves engaging in complex tasks that often lack definitive answers (Biggs & Tang, 2011). This process can be disrupted by the innovation of ChatGPT, as students might prioritise convenience over engagement with challenging learning materials and projects. Therefore, understanding the effect of ChatGPT is crucial for optimising

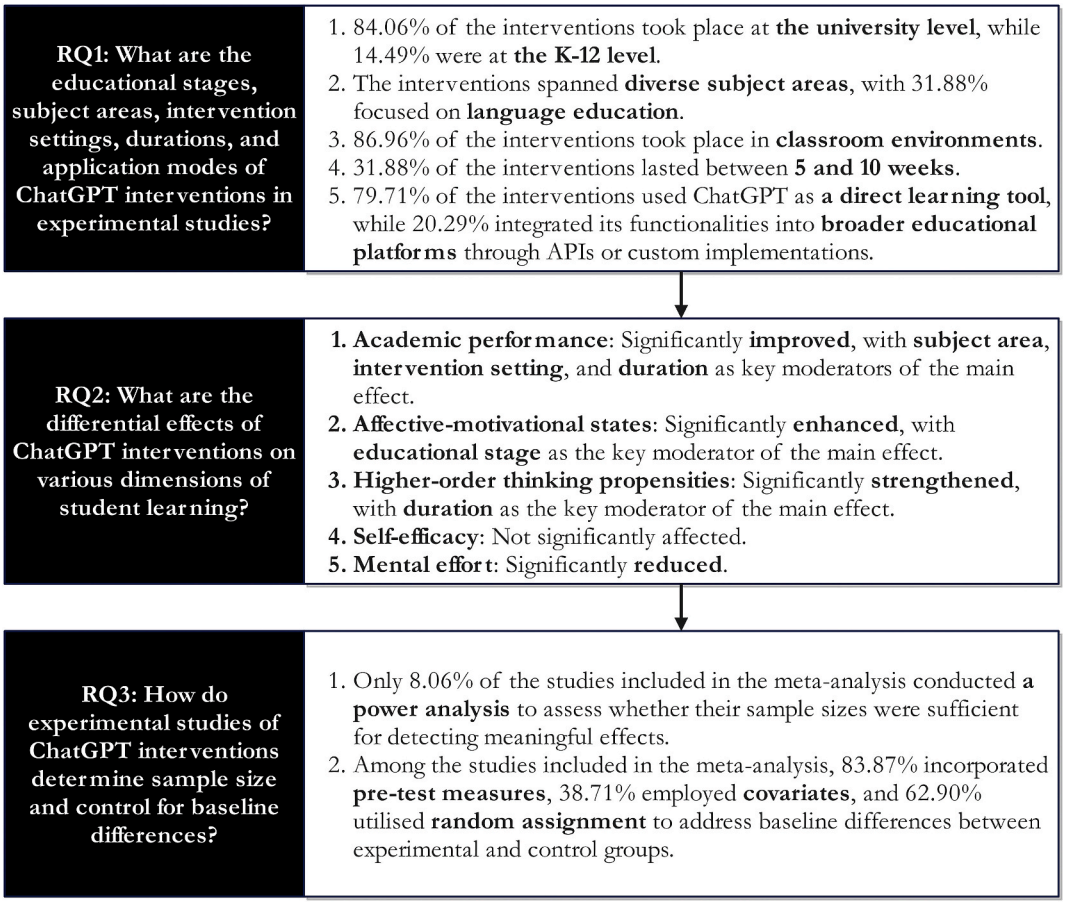


Fig. 2. Key findings by research question.

its integration at the higher education level.

Leading universities often provide publicly accessible guidelines on GenAI use (Dang & Wang, 2024; Moorhouse et al., 2023); however, schools often lack clear policies and administrative support on the matter, and K-12 teachers are calling for school-wide policies that address the ethical use of GenAI (Hays et al., 2024). Empirical evidence is urgently needed to understand the impact of the technology on K-12 learners and to help practitioners formulate effective policies. However, directly applying findings from university-level studies may not accurately reflect the potential benefits or challenges of ChatGPT for younger learners. Boundary conditions such as age could potentially moderate the effectiveness of educational interventions (Deng & Gao, 2023). The extensive focus on university-level education suggests that the potential of ChatGPT in K-12 settings remains underexplored. This trend is evident in existing reviews: a review on ChatGPT in higher education includes 69 studies (Ansari et al., 2024), whereas a similar review on K-12 education covers only 13 studies (Zhang & Tur, 2023). The observation aligns with recent research (Lo et al., 2024; Park & Doo, 2024) and echoes the perspectives of school teachers (Chiu, 2023) and leaders (Dunnigan et al., 2023), highlighting the need for future studies to evaluate the application and effectiveness of ChatGPT among K-12 students to understand its broader impact.

Second, the reviewed studies span diverse subject areas, with a particular emphasis on language education (31.88%). This focus suggests that this field might be particularly well-suited for ChatGPT integration. Previous research supports this notion, indicating that ChatGPT is effective in language education because it can articulate ideas clearly, enhance scientific writing, translate text between languages, and ensure accuracy and grammatical correctness (Lo et al., 2024). This review also highlights the potential of ChatGPT applications in health education, echoing the fact that ChatGPT outperforms search engines in areas such as medical diagnoses (Sandmann et al., 2024). However, the review found that only three studies are conducted in the context of mathematics education, and one study is conducted in the context of legal education. This observation is consistent with findings that ChatGPT performs only satisfactorily or even unsatisfactorily in domains such as mathematics and law (Lo, 2023). This underperformance likely owes to a combination of factors, such as a tendency to invent facts to support conclusions and an overreliance on memorised solutions rather than genuine understanding (Collins et al., 2024). As LLMs evolve, exploring whether updated models can produce more accurate results and better facilitate student learning in these subject areas is critical. Overall, the findings support the idea that ChatGPT's ability to provide clear explanations may be particularly beneficial in certain subject areas (Chiasson et al., 2024). Furthermore, the findings align with previous research suggesting that, despite researchers from various academic disciplines acknowledging the potential of ChatGPT, further empirical research is needed to evaluate its application in under-investigated fields, such as mathematics (Lo et al., 2024).

Third, the reviewed studies are conducted predominantly in classroom settings (86.96%), with a smaller proportion conducted in laboratory environments (8.70%). The predominance of classroom settings suggests that ChatGPT interventions are tested in authentic classroom environments. Despite potential limitations in controlling for confounders and isolating intervention effects, classroom settings provide enhanced ecological validity. Furthermore, intervention durations vary widely from 5 to 10 weeks (31.88%), less than a week (21.74%), 1–4 weeks (17.39%) to more than 10 weeks (14.49%). This variability is consistent with existing research showing that the duration of intervention studies investigating the impact of ChatGPT range from a few lessons or learning tasks to three or four months (Lo et al., 2024). Together, the predominance of classroom settings and the trend towards longer-term interventions suggest a growing recognition among researchers that ChatGPT is not merely a quick fix but an application with the potential for sustained and meaningful impact on educational practices (Korseberg & Elken, 2024; Rawas, 2024).

Finally, most studies (79.71%) use ChatGPT as a direct learning tool, whereas fewer studies integrate ChatGPT's functionalities into broader educational platforms. Researchers compare scenarios using standard ChatGPT applications with conditions where ChatGPT was not used (Wang & Feng, 2023) and with conditions involving alternative technologies like Termbot (Hsu, 2023). This focus reflects their interest in assessing ChatGPT's efficacy as an emerging instructional technology. In contrast, studies that integrate ChatGPT into broader educational platforms (20.29%) are motivated by two primary goals. First, researchers believe that directly using ChatGPT might lead to drawbacks, such as overreliance. To mitigate this, they adapt ChatGPT to offer scaffolding for answering students' questions rather than providing direct answers, aiming to develop higher-order thinking skills such as critical thinking and problem-solving (H.-Y. Lee, Chen, et al., 2024). Second, they seek to explore how integrating ChatGPT into existing tools, such as educational games (Chen & Chang, 2024), could enhance these platforms and create more engaging learning experiences. This observation highlights the importance of not only assessing ChatGPT's effectiveness in direct applications but also investigating innovative uses that leverage its capabilities to foster deeper learning and personalised support.

5.2. RQ2: what are the differential effects of ChatGPT interventions on various dimensions of student learning?

The investigation of RQ2 revealed that ChatGPT interventions significantly improve academic performance, affective-motivational states, and higher-order thinking propensities. The results also demonstrate that ChatGPT interventions significantly reduce mental effort but have no significant effect on self-efficacy. The following sections discuss these key findings.

5.2.1. Academic performance

The meta-analysis revealed that the adoption of ChatGPT in education significantly improved students' academic performance compared to non-adoption. ChatGPT may genuinely enhance learning performance by enabling personalised learning experiences

(Wang et al., 2024), providing immediate access to information and diverse perspectives (Urban et al., 2024), and allowing students to engage more deeply with material (Meyer et al., 2024). However, the promising results warrant cautious interpretation. An alternative explanation for the improved academic performance could be the higher quality of work produced with ChatGPT's assistance, which might be misconstrued as genuine improvement in student learning. A closer examination of the meta-analysed studies reveals that nine permit participants to use ChatGPT during post-intervention assessments of academic performance (Bašić et al., 2023; Li, 2023), and 33 do not explicitly state whether ChatGPT was permitted during academic performance assessments (e.g. Alneyadi & Wardat, 2023; Song & Song, 2023). This methodological detail is crucial because the observed positive effects could be attributed to the inherent quality of ChatGPT-generated content rather than the intervention itself. In-class, proctored assessments could be adopted to mitigate GenAI-assisted plagiarism (Chaudhry et al., 2023; Newton & Xiromeriti, 2024) and ensure a rigorous evaluation of intervention effectiveness. Future research should explicitly state whether ChatGPT is accessible to learners during post-intervention assessments (e.g. Stadler et al., 2024; Urban et al., 2024) and, if applicable, describe how learners use it.

When the adoption of ChatGPT during assessments is permitted or unavoidable, researchers can employ strategies to differentiate between the quality of ChatGPT's output and the outcomes of educational interventions. One approach is to shift from well-defined problems (Kim et al., 2024) and cognitive-intensive tasks (Zhai et al., 2024) to project-based assessments (Liu, 2024) that require students to demonstrate practical application, knowledge integration, and a broader range of analytical and creative skills. Such assessments often involve unique contexts and personal experiences that would be challenging for ChatGPT to replicate authentically. Assessment genres that outpace ChatGPT's capabilities, such as requiring oral components as part of an assessment (Wise et al., 2024), can also be considered. These strategies not only distinguish between the assistance provided by GenAI and authentic learning gains, but may also reduce student reliance on ChatGPT-generated content (Perkins et al., 2023) and incentivise genuine cognitive engagement (Waltzer et al., 2024).

This review also highlights the importance of considering metrics beyond quality when evaluating the impact of ChatGPT on academic performance, particularly in writing. Within language education, the reviewed studies typically evaluate writing performance based on multiple subdimensions such as coherence, grammar, and lexical range (Boudouaia et al., 2024), primarily focusing on assessing the quality of writing outputs. Despite evidence supporting the positive impact of ChatGPT on writing (e.g. Mahapatra, 2024), it is important to avoid uncritical acceptance. The study by Niloy et al. (2023), which finds a detrimental effect of ChatGPT, uniquely considers originality when assessing writing performance. This observation resonates with concerns expressed by instructors (Gammoh, 2024; Pack & Maloney, 2024) and students (Karkoulian et al., 2024; Zhao et al., 2024) regarding the potential of ChatGPT to facilitate plagiarism. Furthermore, GenAI-assisted revisions may create a misleading impression of a learner's writing proficiency, as such work may no longer reflect their authentic writing competence (Tsai et al., 2024). Given that ChatGPT can produce higher-quality outputs (de Winter 2023; Vázquez-Cano et al., 2023), such as essays (Herbold et al., 2023), compared to students, solely attributing quality as an indicator of positive impact may inadvertently overshadow potential weaknesses associated with adopting this technology. Incorporating metrics such as originality or authenticity (Higgs & Stornaiuolo, 2024; Ironsi & Ironsi, 2024) alongside quality assessments of student output would provide a more thorough evaluation and gain a deeper understanding of the multifaceted impact of ChatGPT.

The moderator analysis showed that subject area, intervention setting, and duration were significant moderators of the effect of ChatGPT interventions on academic performance. Specifically, interventions in the arts and humanities were associated with a larger effect size, followed by those in health and medical sciences, social sciences, and science. The larger effect size in the arts and humanities, predominately comprising language education, may be attributed to ChatGPT's inherent strengths in natural language processing and text generation. These capabilities align well with core aspects of language learning, such as vocabulary acquisition, grammar practice, and developing writing skills through interactive feedback and text exposure (Karataş et al., 2024; Ma et al., 2024). The moderator analysis also indicated that classroom-based interventions exhibited a significant large effect size, while laboratory settings yielded a non-significant effect. Furthermore, interventions lasting 1–4 weeks showed a larger effect size compared to those lasting 5–10 weeks and over 10 weeks, while interventions under one week demonstrated no significant effect. Interventions in laboratory settings and those lasting less than one week demonstrated limited efficacy, possibly because they did not reflect real-world learning contexts and were less ecologically valid. Laboratory environments often lack the contextual relevance, student motivation, and social dynamics found in classrooms (Anderson & Shattuck, 2012; Deng & Gao, 2023), which can potentially diminish the impact of ChatGPT interventions. Similarly, brief interventions may not provide sufficient time for students to fully engage with and benefit from ChatGPT (Çelik, Ersanlı, & Arslanbay, 2024), resulting in limited improvements in academic performance. While these results suggest that certain subjects may be more suitable for ChatGPT integration and that the effects of the interventions may be associated with intervention setting and duration, it is crucial to highlight that most reviewed studies did not explicitly state whether ChatGPT was permitted during post-intervention assessments. This implies that the quality of ChatGPT's output is not distinguished from the effect of the intervention itself. Therefore, at this stage, this review refrains from making generalisations about the moderating effects of subject area, intervention setting, and duration on ChatGPT's impact on academic performance. Future research should expand the scope of this review and re-evaluate these moderating effects, particularly as more experiments differentiate between the quality of ChatGPT output and the effects of the interventions.

5.2.2. Affective-motivational states

The meta-analysis indicated that the utilisation of ChatGPT in education had a significant positive impact on students' affective-motivational states. This observation aligns with recent research indicating that adopting ChatGPT tends to elicit positive affective responses (Koltovskaia et al., 2024; Lo et al., 2024; Woo et al., 2024). The positive impact suggests that ChatGPT has the potential to enhance study motivation and make the learning process more enjoyable. However, it is important to note that among the 16 studies reporting positive effects, 13 measure motivational and emotional factors at two points: before and after the intervention (e.g. Silitonga et al., 2023). Concurrently, three assess these factors solely post-intervention (e.g. Wu et al., 2024). Despite this review showing that ChatGPT interventions typically extend over several weeks rather than being one-off randomised control trials (detailed in Section 5.1), none of the reviewed studies measure participants' affective-motivational states multiple times throughout these interventions. This means the potential fluctuation of affective-motivational states over time is overlooked. Qualitative evidence indicates that learners have mixed opinions about the effectiveness of ChatGPT in fostering motivation or providing emotional support (Rienties et al., 2024). While this review suggests that ChatGPT may have a promising impact on students' affective-motivational states, the observed effects might have been influenced by a novelty effect, where the initial excitement of interacting with a new technology could have temporarily inflated their positive perceptions (Zhai & Wibowo, 2023). Future research should investigate these fluctuations (Croes & Antheunis, 2021) and the longer-term impact of ChatGPT (Polyportis, 2024) on affective-motivational states to ascertain whether these positive effects are sustained over time or merely attributable to the novelty associated with adopting the technology.

The moderator analysis revealed that ChatGPT interventions in university settings demonstrated a large effect size, whilst those in K-12 settings showed no significant effect. This finding implies that the impact of ChatGPT interventions on affective-motivational states is more pronounced for college-aged learners compared to younger students. A possible explanation is that the complexity of tasks in higher education aligns better with the capabilities of ChatGPT (Baig & Yadegaridehkordi, 2024). Additionally, university students may have more experience with novel technologies, making them more comfortable utilising GenAI tools for academic support (Ansari et al., 2024). The results warrant further investigation to understand why the impacts of ChatGPT interventions on affective-motivational states vary across different educational stages.

5.2.3. Higher-order thinking propensities

The meta-analysis demonstrated that the integration of ChatGPT in education significantly enhanced students' higher-order thinking propensities. While some scholars emphasise the potential of ChatGPT in cultivating higher-order thinking skills (van den Berg & du Plessis, 2023), others question the tool's effectiveness in fostering these skills (Yang & Li, 2024). The improvements in higher-order thinking propensities appears to highlight the potential of ChatGPT to support complex cognitive processes and alleviate concerns that ChatGPT could negatively impact the development of higher-level thinking (Valcea et al., 2024). However, these studies focus on students' self-appraisals of higher-order thinking rather than their actual, demonstrated skill acquisition and development. Sole reliance on subjective measurement of propensity, tendency, awareness, or disposition alone may not accurately capture the impact. Experimental evidence shows that students may perceive they have learned less when, in fact, they have learned more, and conversely, they may feel they have learned more even when their actual learning is less (Deslauriers et al., 2019). Individuals might overestimate their higher-order thinking propensities for various reasons, such as social desirability bias, where they align with perceived expectations (Paulhus, 1991), or self-enhancement bias, where they seek to improve their self-image (Kwan et al., 2008). To address these potential limitations, future research could triangulate self-report data with objective measures, such as standardised tests that assess critical thinking (Roohr et al., 2019) or problem-solving tasks that require students to actively demonstrate their skills (Kapur et al., 2023), to offer a more comprehensive understanding of the true effects of ChatGPT interventions on higher-order thinking.

The moderator analysis indicated that ChatGPT interventions with shorter durations generally yielded a larger effect size. This suggests that the effectiveness of ChatGPT interventions on higher-order thinking propensities may decrease with prolonged use. Alternatively, it is possible that individuals may become more adept at calibrating their higher-order thinking beliefs as educational interventions extend over time (Dunning et al., 2003; Veenman et al., 2006). Future research should investigate why intervention duration moderates the impact of ChatGPT on higher-order thinking propensities.

5.2.4. Self-efficacy

The meta-analysis indicated that the implementation of ChatGPT in education had a non-significant effect on students' self-efficacy. ChatGPT interventions enhance self-efficacy in some studies (e.g. Li, 2023; Urban et al., 2024) but not in others (e.g. Aydın Yıldız, 2023; Donald et al., 2024). Possible reasons for the non-significant effect size include the diverse approaches in which ChatGPT is integrated into the learning process for experimental groups and the context-specific, task-dependent nature of self-efficacy (Bandura, 1997). Although there is heterogeneity among the reviewed studies, conducting moderator analyses with a small number of studies can yield unreliable results (Baker et al., 2018). Instead of drawing definitive conclusions suggesting that ChatGPT has no impact on students' self-efficacy, this review encourages future meta-analyses to re-evaluate this relationship as more experimental studies become available.

Although discrete studies have reported positive effect of ChatGPT on self-efficacy, they have not thoroughly explored the

mechanisms through which ChatGPT improves self-efficacy. Bandura (1997) proposes that key predictors of self-efficacy are mastery experiences, vicarious experiences, social persuasion, and emotional and physiological states. It is possible that these improvements were achieved through mastery experiences (e.g. providing tailored, individualised feedback), as qualitative evidence from Li (2023) indicates that ChatGPT provided learners with opportunities to successfully complete tasks they might have found challenging otherwise. Moreover, the positive impact of ChatGPT on learners' affective-motivational states found in this review suggests that enhanced emotional and physiological states may also play a role (e.g. offering empathetic responses and encouragement). Future research should investigate why ChatGPT positively enhances self-efficacy in specific teaching and learning contexts while not in others, as well as how this effect occurs.

5.2.5. Mental effort

The meta-analysis showed that the application of ChatGPT in education significantly reduced students' mental effort, making learning less cognitively demanding. Mental effort reflects the amount of cognitive resource that an individual exerts to perform a learning task and is often used as an indicator of cognitive load due to its relative ease of measurement (Krieglstein et al., 2022; Mutlu-Bayraktar et al., 2019). Past research shows that students can experience heavy mental effort when writing with ChatGPT (Woo et al., 2024), but the cross-sectional nature of that study makes it challenging to determine if this owes to the novelty of the technology, the demanding writing task, or the time constraints of the writing workshop. Experimental studies that manipulate ChatGPT while controlling for other factors can address this limitation. Based on such experimental evidence (e.g. Ji et al., 2023), this review demonstrates that ChatGPT integration could reduce mental effort compared to conditions without it, highlighting the importance of experimental research in isolating factors that may influence mental effort.

The effectiveness of an instructional condition is deemed high if learners can achieve strong performance with minimal mental effort and considered low if significant mental effort results in poor performance (Paas et al., 2005). A closer examination of the studies that report positive effects on mental effort revealed concurrent improvements in academic performance. This appears to reinforce the view that the instructional condition that integrates ChatGPT benefits mental effort and learning performance. However, caution is warranted as two studies allowed the use of ChatGPT during the post-test (Ji et al., 2023; Urban et al., 2024), and one study did not disclose this information (T. Li, Ma, et al., 2024). Notably, one study explicitly prohibits the use of ChatGPT during the post-test, revealing that students who use ChatGPT demonstrate reduced mental effort; however, they simultaneously exhibit weaker reasoning and argumentation in the subsequent test (Stadler et al., 2024). This observation not only highlights the potential for reduced mental effort to come at the expense of deeper learning but also underscores the importance of scholars explicitly stating whether ChatGPT was available to learners during post-tests to disentangle the quality of ChatGPT output from the effect of educational interventions.

5.3. RQ3: how do experimental studies of ChatGPT interventions determine sample size and control for baseline differences?

The exploration of RQ3 revealed a notable absence of power analysis to determine adequate sample sizes across the studies. Despite this, most studies employ random assignment, pre-tests, or covariates to address baseline differences between experimental and control groups. Subsequent sections discussed key findings.

First, the sample sizes in the meta-analysed studies range from a mere 18 to a substantial 600 participants, and only 8.06% of these studies report conducting a power analysis to ensure their sample sizes were sufficient to detect meaningful effects. This lack of power analysis raises concerns about the validity of the findings, as underpowered studies are prone to Type II errors, where a true effect is missed owing to insufficient sample size (Abu-Bader, 2021; Sommet et al., 2023). For example, Bašić et al. (2023) with only 18 participants, may have failed to detect a statistically significant effect of ChatGPT on writing performance, even if such an effect existed. This false negative outcome could have significant implications, potentially dissuading researchers from further exploring the potential benefits of ChatGPT for writing. Furthermore, underpowered studies may produce unreliable effect size estimates (Schmidt et al., 2018; Sommet et al., 2023). For instance, Wiboolyasarin et al. (2024) find a significant positive effect of ChatGPT on writing performance with a sample size of 39; however, the absence of a power analysis raises the question of whether the observed effect size accurately reflects the true impact. If the study was underpowered, the reported improvement in motivation could be larger than the actual effect in the population (Cohen, 1994). This could mislead education researchers and practitioners into overestimating the benefits of ChatGPT and investing resources in interventions that may not yield the expected results in practice. The observation highlights the need for future research to prioritise adequate sample sizes and power analysis to ensure the validity of findings. By doing so, scholars can provide more accurate and trustworthy insights into the effects of ChatGPT.

Second, more than 83% of the reviewed studies employ random assignment, pre-tests, control for covariates, or a combination of these methods to address baseline differences, reflecting the prevalent effort to ensure group comparability. Randomisation is a key method for controlling baseline differences—by randomly assigning participants to different intervention groups, it balances both known and unknown variables, thereby reducing the impact of confounders (Sterne et al., 2019). This review showed that randomisation was used in 39 (62.90%) of the meta-analysed studies. Notably, among these 39 studies, 16 (41.03%) conduct pre-tests and use pre-test scores as covariates alongside random assignment. This combination reflects a more robust approach that enhances internal validity by addressing baseline differences through pre-tests and adjusting for them statistically with covariates, thereby increasing

confidence that observed post-intervention differences are attributable to the intervention rather than pre-existing differences (Shadish et al., 2002).

However, random assignment is often hindered by factors such as cost, time constraints, and ethical considerations (Lee et al., 2017), making it less feasible in educational contexts where students are already grouped into classroom cohorts. When randomisation is not feasible, researchers may employ pre-tests to evaluate baseline differences, and if inequivalence is found, use pre-test data as covariates to statistically adjust for baseline differences (Sterne et al., 2019). This review found that seven studies (11.29%) opt not to use random assignment but conduct pre-tests and use pre-test scores as covariates, indicating an effort to control for pre-existing differences between groups. Despite this, such approaches may struggle to rule out alternative explanations (Shadish et al., 2002). This is because pre-tests and covariates can only control for known differences between groups and cannot account for all potential confounding variables, particularly those that are not considered in the experimental design. For instance, factors such as motivation, attitude, and prior ChatGPT experience could influence academic performance but might not be adequately captured by prior knowledge test scores (Hsu, 2023; Mahapatra, 2024). Consequently, there remains the possibility that observed differences are driven by these unmeasured factors rather than the ChatGPT intervention. It is advisable to enhance quasi-experimental designs by, for example, incorporating a more extensive set of baseline measures, developing advanced matching techniques to create more comparable groups, and utilising multiple control groups to account for different potential confounders (Shadish et al., 2002), thus improving the isolation of intervention effects in settings where random assignment is impractical.

6. Research implications

The findings of this review highlight several key implications for education researchers investigating the impact of ChatGPT and similar GenAI tools on student learning through experimental studies. First, researchers should determine whether the positive effects of ChatGPT interventions on academic performance are owing to the inherent quality of ChatGPT response or the collaborative role of ChatGPT. This can be achieved by employing strategies such as shifting from well-defined problems to more complex, project-based assessments that require demonstration of a range of skills, adopting proctored assessments, and incorporating metrics (e.g. originality) alongside assessments of quality. Clarifying this distinction is crucial because many studies allow participants to use ChatGPT during post-tests, which makes it challenging to discern whether observed positive effects stem from the high output quality of ChatGPT or the impact of interventions.

Second, researchers should consider evaluating long-term impacts and providing insights into whether the positive effects of ChatGPT on students' affective and motivational states are sustained over time or merely a function of initial exposure to the technology. For instance, experimental studies could track students' affective and motivational levels over multiple months or semesters to provide practical insights into ChatGPT's role in fostering positive emotions and sustained motivation.

Third, researchers should supplement students' self-appraisals of higher-order thinking propensities, tendencies, awareness, or dispositions with objective measures. For instance, they could create authentic learning scenarios that require students to apply higher-order thinking without ChatGPT assistance and then compare these outcomes to tasks completed with ChatGPT support to assess genuine skill development. Simultaneously, students could be asked to subjectively evaluate their perceived improvement in higher-order thinking. This approach would help address key questions, such as whether ChatGPT facilitates actual gains in higher-order thinking skills or merely enhances students' perceptions of skill development.

Finally, researchers should conduct power analyses to ensure sufficient sample sizes, which would reduce the risk of Type II errors and yield reliable effect sizes. In addition, employing methods such as random assignment, pre-tests, and covariates to control for baseline differences would enhance the robustness of future studies. This would help ensure that the effects observed are truly due to the intervention and not confounded by pre-existing conditions.

7. Limitations and future directions

This review identifies several opportunities to advance the field, but four caveats related to the study's approach must be considered. First, this review directly examined the impact of ChatGPT as a novel instructional technology on student learning, without situating this impact within a specific learning theory or framework. This approach was adopted because, while 41 of the 69 reviewed studies are atheoretical, others draw on a variety of theories. The most frequently mentioned theory is constructivist learning theory ($n = 7$), followed by self-regulated learning ($n = 4$), the technology acceptance model ($n = 3$), and cognitive load theory ($n = 3$). Social cognitive theory, scaffolding theory, experiential learning theory, and control-value theory each appeared twice in the reviewed studies. Additionally, other theories such as self-determination theory, flow theory, and distributed learning theory were each mentioned once. Therefore, adopting a unified theoretical framework to structure this review would have been challenging. Although we could have chosen to focus on experimental studies that apply a specific theory, such as cognitive load theory, doing so would have significantly restricted the scope of the review. Instead, this review followed the conventions of existing meta-analyses (e.g. Schroeder et al., 2023), taking a pragmatic approach to investigating the impact of a selected innovative instructional technology on student learning. As research in this field becomes more mature and sophisticated, it is recommended that future systematic reviews and meta-analyses incorporate a specific theoretical lens to better understand how ChatGPT influences learning-related outcomes.

Second, categorising outcome variables such as self-efficacy (Bandura, 1982) and mental effort (Sweller et al., 1998) helps understand the impact of ChatGPT on student learning; however, these constructs originate from different theoretical frameworks and have diverse theoretical foundations. Juxtaposing them might introduce ontological constraints, and there are undoubtedly alternative ways to organise learning outcome variables. Future research should develop a more cohesive conceptual framework that integrates these diverse constructs to provide a clearer understanding of the multifaceted impact of ChatGPT on learning. Exploring alternative classifications of learning outcomes can help refine the evaluation of ChatGPT effectiveness in educational contexts.

A third limitation is that non-English publications were excluded from the article selection process due to the researchers' language expertise. Such studies may provide valuable insight into the effectiveness of ChatGPT idiosyncratic to learners whose primary language is not English. Future research incorporating publications written in other languages can improve the understanding of ChatGPT's impact on students from diverse linguistic backgrounds.

Finally, this review focused exclusively on the effects of ChatGPT on student learning, without considering other GenAI products. While ChatGPT is widely used, other models may offer unique features that benefit teaching and learning in certain subjects. Future research could explore the effectiveness of various GenAI tools and provide a more comprehensive understanding of how different models impact student learning across diverse instructional contexts and student populations.

8. Conclusions

Early research on ChatGPT in educational settings focuses on the perceptions and attitudes of both students (Lee & Zhai, 2024) and instructors (Cambra-Fierro et al., 2024) towards the technology. These studies reveal that students generally exhibit positive attitudes towards ChatGPT (Dube et al., 2024; Haindl & Weinberger, 2024), whereas instructors hold more ambivalent views (Al-khresheh, 2024; Derakhshan & Ghiasvand, 2024). However, perceptions or attitudes alone do not provide concrete evidence of the actual impact of ChatGPT on learning. Moreover, cross-sectional research has shown both positive (Shahzad et al., 2024) and negative associations (Crawford et al., 2024) between ChatGPT usage and academic performance. However, these studies cannot determine whether ChatGPT use leads to improved or diminished performance, or vice versa. These limitations underscore the need for experimental studies to ascertain the impact of ChatGPT on student learning. Existing reviews of ChatGPT in education (e.g. Samala et al., 2024) have offered insights into the potential ways in which teaching and learning can be affected by ChatGPT. While acknowledging their value, these reviews do not capture the technology's actual impact on learning outcomes. To date, no systematic review has been conducted to examine the impact of ChatGPT on learning or to reconcile the differential impacts observed in discrete experimental studies.

To address the knowledge gap, this study conducted a systematic review and meta-analysis to synthesise research findings on the impact of ChatGPT interventions on student learning. This review makes a significant contribution to the field of technology-enhanced learning by illuminating the general characteristics of experimental studies that explore the effects of ChatGPT on various dimensions of student learning, clarifying the differing impacts observed across these studies, and scrutinising their methodological quality. Specifically, it revealed that ChatGPT interventions are predominantly concentrated at the university level, span a wide range of subject areas with a particular emphasis on language education, are typically integrated into authentic classroom environments as part of regular educational practices rather than short-term isolated tools, and are mostly used as direct learning tools with a smaller proportion being incorporated into broader educational platforms. Most meta-analysed studies use random assignments, pre-tests, control for covariates, or a combination of these methods to address baseline differences, demonstrating an effort to manage pre-existing differences between groups. However, this review identified a critical limitation in the infrequent use of power analysis for determining sufficient sample sizes, which may lead to Type II errors and unreliable estimates of effect size.

Notably, the review suggests that ChatGPT can potentially improve academic performance, as evidenced by the overall large, positive effect. Simultaneously, it highlights the need for caution in interpreting these results owing to limitations in methodological approaches (e.g. power analysis deficiencies) and assessment concerns (e.g. allowing the use of ChatGPT during post-intervention assessments). In addition, the review indicates that ChatGPT interventions enhanced affective-motivational states and higher-order thinking propensities while reducing mental effort. Nevertheless, there is a general lack of experimental research investigating the fluctuations or long-term impacts of ChatGPT to rule out the novelty effect of the technology, or using objective measures to evaluate actual higher-order thinking skills. These findings provide important implications for future interventions aimed at investigating the impact of ChatGPT. Future research is advised to employ strategies to disentangle the quality of ChatGPT output from the intervention effects, explore fluctuations and the long-term impact of ChatGPT on affective-motivational states to determine whether the observed benefits are due to a novelty effect, and assess whether improvements in students' self-appraisals of higher-order thinking correspond to actual gains in these skills.

CRediT authorship contribution statement

Ruiqi Deng: Writing – original draft, Writing – review & editing, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Maoli Jiang:** Validation, Data curation. **Xinlu Yu:** Validation, Data curation. **Yuyan Lu:** Validation, Data curation. **Shasha Liu:** Writing – review & editing, Validation, Methodology, Conceptualization.

Funding

This study was funded by the National Natural Science Foundation of China [Grant number 72204072] and the Zhejiang Province Education Science Planning Project [Grant number 2024SCG333].

Declaration of competing interest

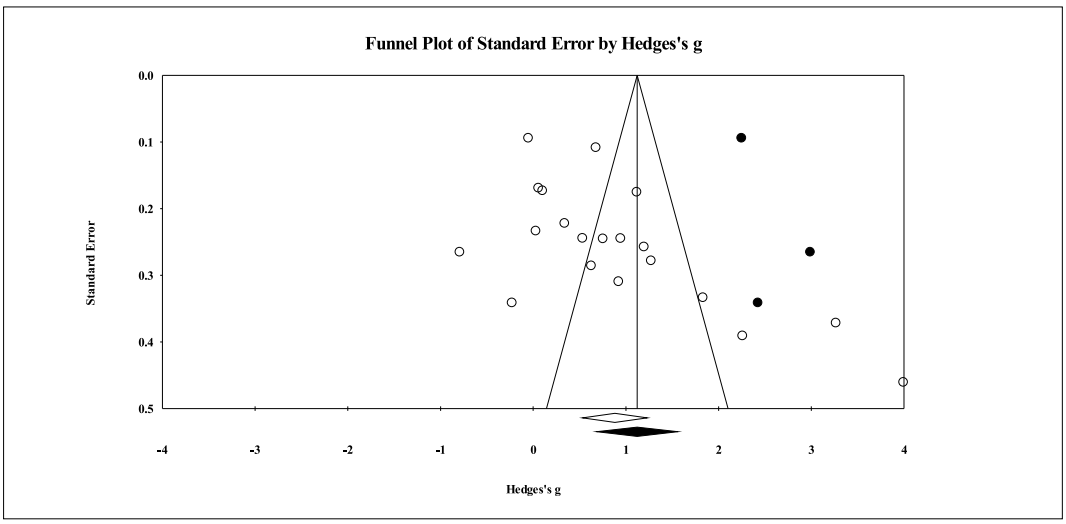
The authors declare that they have no competing interests.

Acknowledgements

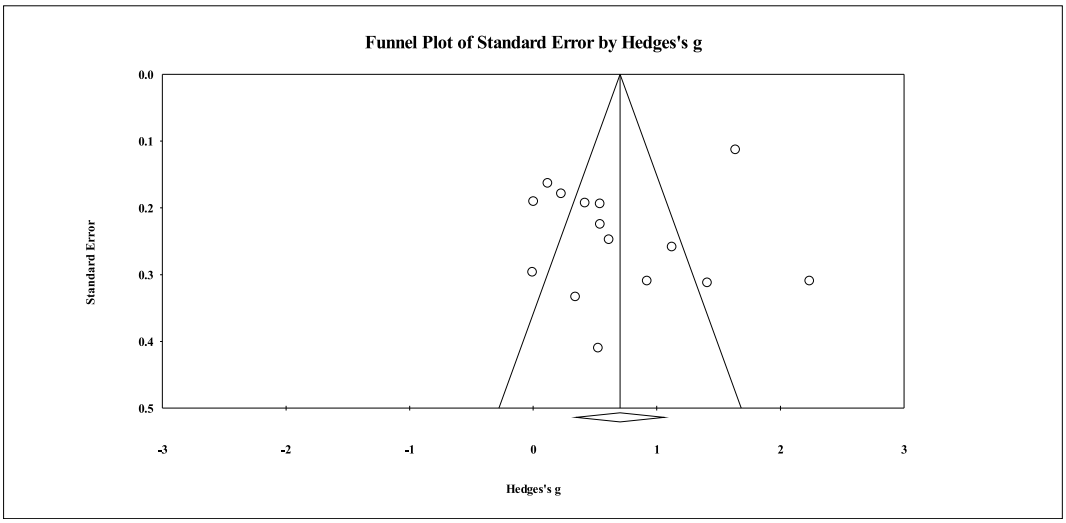
We thank the three anonymous reviewers for their constructive feedback, Dr Jinbo He and Dr Suqin Shen for their insightful comments on earlier drafts of this article, and Ms Ziluo Zhang for her valuable assistance with data curation.

Appendix A. Funnel plot of outcome measures

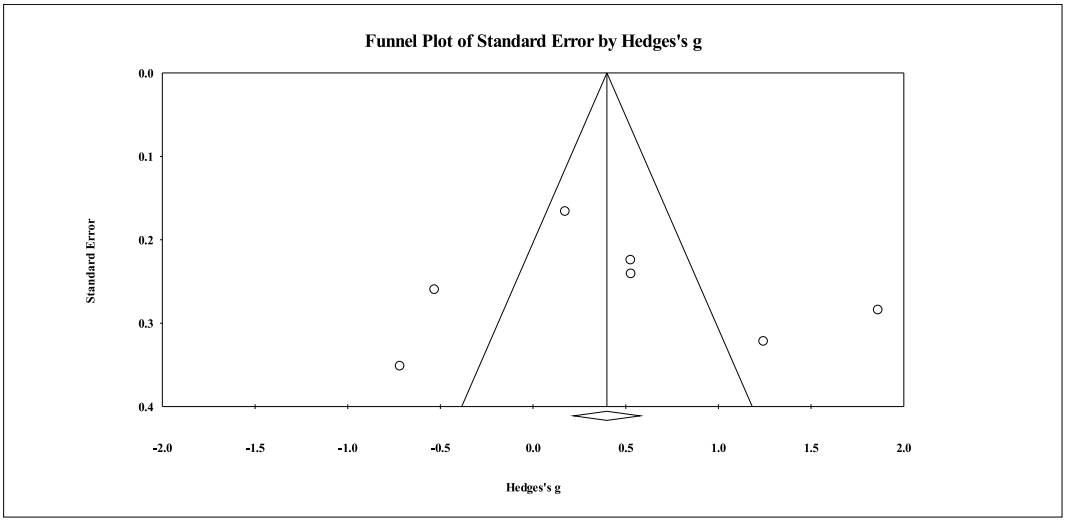
Appendix A1. Funnel plot of effect sizes for academic performance



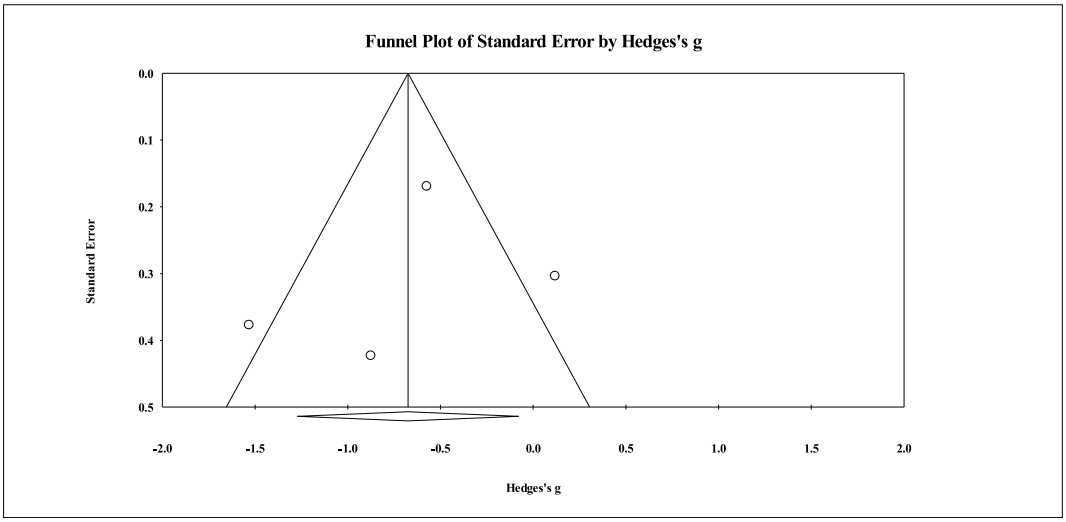
Appendix A2. Funnel plot of effect sizes for affective-motivational states.



Appendix A3. Funnel plot of effect sizes for higher-order thinking propensities.



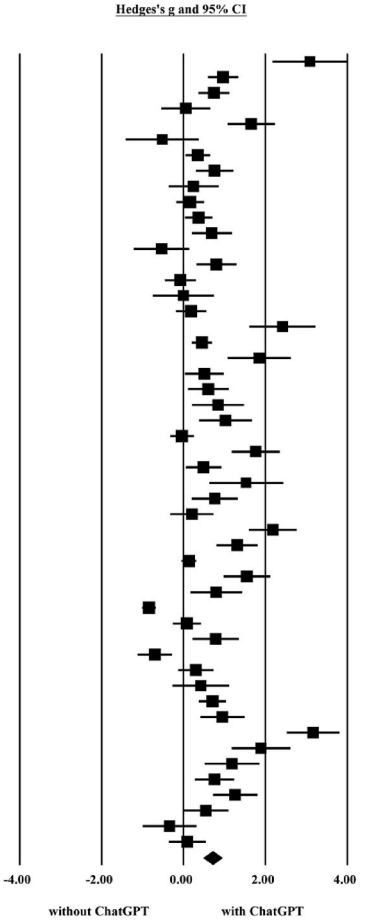
Appendix A4. Funnel plot of effect sizes for self-efficacy.



Appendix A5. Funnel plot of effect sizes for mental effort.

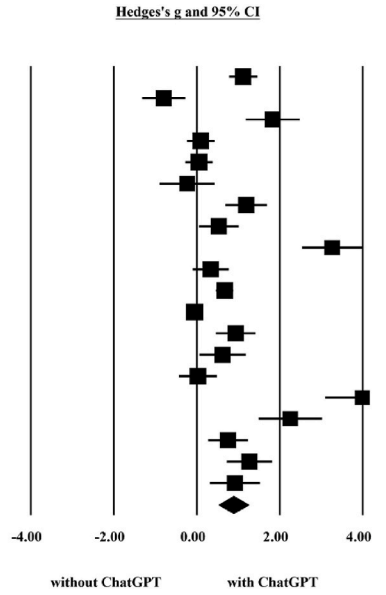
Appendix B. Forest plots of effect sizes for outcomes measures

Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Ahmed Moneus and Al-Wasy (2024)	3.089	0.464	0.215	2.179	3.998	6.656	0.000
Alneyadi and Wardat (2023)	0.973	0.191	0.036	0.600	1.347	5.105	0.000
Alneyadi and Wardat (2024)	0.748	0.194	0.038	0.367	1.129	3.852	0.000
Avello-Martinez, Gajderowicz and Gómez-Rodríguez (2024)	0.062	0.306	0.094	-0.539	0.663	0.202	0.840
Aydın Yıldız (2023)	1.658	0.296	0.088	1.077	2.239	5.594	0.000
Bašić et al. (2023)	-0.516	0.457	0.209	-1.412	0.380	-1.128	0.259
Beltozar-Clemente and Díaz-Vega (2024)	0.352	0.154	0.024	0.050	0.654	2.287	0.022
Boudouaia, Mouas and Kouider (2024)	0.762	0.235	0.055	0.301	1.224	3.237	0.001
Chandrasekera, Hosseini and Perera (2024)	0.251	0.311	0.097	-0.359	0.861	0.805	0.421
Chen and Chang (2024) - Comp.1	0.169	0.173	0.030	-0.170	0.508	0.975	0.330
Chen and Chang (2024) - Comp.2	0.374	0.170	0.029	0.041	0.708	2.200	0.028
Darmawansah et al. (2024)	0.697	0.249	0.062	0.210	1.185	2.802	0.005
Donald, Will and Christopher (2024)	-0.530	0.346	0.120	-1.208	0.148	-1.531	0.126
Emran et al. (2024)	0.809	0.250	0.062	0.319	1.298	3.240	0.001
Essien et al. (2024)	-0.072	0.192	0.037	-0.449	0.305	-0.375	0.708
Farah et al. (2023)	0.004	0.380	0.144	-0.740	0.749	0.011	0.991
Gan et al. (2024)	0.193	0.190	0.036	-0.179	0.565	1.018	0.309
Hsu (2023)	2.420	0.411	0.169	1.613	3.226	5.882	0.000
Huesca et al. (2024)	0.453	0.127	0.016	0.203	0.702	3.555	0.000
Ji et al. (2023)	1.854	0.396	0.157	1.078	2.631	4.680	0.000
Kavadella et al. (2024)	0.514	0.242	0.059	0.040	0.988	2.126	0.034
Khuibut, Premthaisong and Chaipidech (2023)	0.612	0.253	0.064	0.115	1.108	2.415	0.016
Kim (2024).Comp1	0.849	0.324	0.105	0.214	1.484	2.620	0.009
Kim (2024).Comp2	1.034	0.331	0.109	0.386	1.683	3.126	0.002
Kosar et al. (2024)	-0.029	0.148	0.022	-0.318	0.260	-0.196	0.844
Kucuk (2024)	1.770	0.302	0.091	1.178	2.361	5.865	0.000
Li (2023)	0.493	0.224	0.050	0.055	0.932	2.206	0.027
T. Li et al. (2024)	1.537	0.462	0.213	0.632	2.442	3.328	0.001
Lyu et al. (2024)	0.768	0.289	0.083	0.202	1.334	2.659	0.008
Maghamil and Sieras (2024)	0.212	0.269	0.072	-0.315	0.739	0.788	0.430
Mahapatra (2024).Comp1	2.186	0.296	0.088	1.606	2.766	7.386	0.000
Mahapatra (2024).Comp2	1.316	0.258	0.066	0.811	1.822	5.107	0.000
Meyer et al. (2024)	0.133	0.094	0.009	-0.051	0.317	1.412	0.158
Mugableh (2024)	1.553	0.292	0.085	0.982	2.125	5.326	0.000
Mun (2024)	0.807	0.323	0.104	0.174	1.439	2.499	0.012
Niloy et al. (2023)	-0.836	0.085	0.007	-1.003	-0.670	-9.834	0.000
Shi, Li and Zhang (2024)	0.089	0.176	0.031	-0.256	0.434	0.507	0.612
Song and Song (2023)	0.790	0.289	0.084	0.223	1.357	2.729	0.006
Stadler, Bannert and Sailer (2024)	-0.696	0.214	0.046	-1.116	-0.276	-3.246	0.001
Sun et al. (2024)	0.306	0.220	0.049	-0.125	0.738	1.391	0.164
Svendsen et al. (2024)	0.429	0.354	0.125	-0.266	1.123	1.210	0.226
Urban et al. (2024)	0.713	0.171	0.029	0.379	1.048	4.177	0.000
Wahba, Ajlouni and Abumosa (2024)	0.955	0.279	0.078	0.409	1.501	3.430	0.001
Wang and Feng (2023)	3.168	0.328	0.108	2.524	3.811	9.650	0.000
Wang et al. (2024)	1.899	0.367	0.135	1.179	2.619	5.172	0.000
Wiboolyasarin et al. (2024)	1.190	0.342	0.117	0.519	1.861	3.476	0.001
Wu et al. (2023)	0.763	0.245	0.060	0.283	1.243	3.114	0.002
Wu et al. (2024)	1.266	0.278	0.077	0.722	1.811	4.562	0.000
Xiao (2024)	0.553	0.281	0.079	0.002	1.104	1.965	0.049
Zhang et al. (2024)	-0.338	0.337	0.114	-0.999	0.324	-1.001	0.317
Zhou and Kim (2024)	0.096	0.230	0.053	-0.355	0.548	0.418	0.676
	0.712	0.109	0.012	0.497	0.926	6.502	0.000



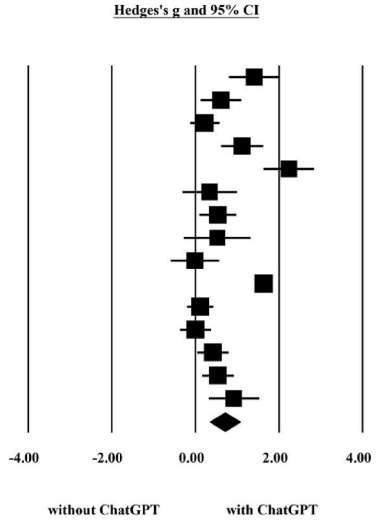
Appendix B1. Forest plots of effect sizes for academic performance.

Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Almohesh (2024)	1.117	0.175	0.031	0.774	1.460	6.390	0.000
Aydın Yıldız (2023)	-0.794	0.265	0.070	-1.313	-0.274	-2.996	0.003
Chang et al. (2024)	1.831	0.333	0.111	1.178	2.484	5.495	0.000
Chen and Chang (2024) - Comp.1	0.100	0.173	0.030	-0.239	0.438	0.576	0.565
Chen and Chang (2024) - Comp.2	0.056	0.169	0.028	-0.275	0.387	0.333	0.739
Donald, Will and Christopher (2024)	-0.230	0.341	0.116	-0.898	0.439	-0.674	0.501
Hu (2024a)	1.195	0.257	0.066	0.691	1.698	4.647	0.000
Hu (2024b) - Comp.1	0.532	0.244	0.060	0.053	1.010	2.179	0.029
Hu (2024b) - Comp.2	3.265	0.371	0.138	2.537	3.992	8.792	0.000
Li (2023)	0.339	0.222	0.049	-0.096	0.773	1.526	0.127
Y. Li et al. (2024)	0.676	0.108	0.012	0.464	0.888	6.252	0.000
Meyer et al. (2024)	-0.053	0.094	0.009	-0.237	0.131	-0.564	0.573
Silitonga et al. (2023)	0.943	0.244	0.060	0.464	1.422	3.857	0.000
Song and Song (2023)	0.626	0.285	0.081	0.067	1.186	2.195	0.028
Suciati et al. (2024)	0.028	0.233	0.054	-0.429	0.485	0.118	0.906
Wahba, Ajlouni and Abumosa (2024)	3.993	0.460	0.212	3.091	4.895	8.676	0.000
Wang et al. (2024)	2.257	0.391	0.153	1.492	3.023	5.780	0.000
Wu et al. (2023)	0.752	0.245	0.060	0.272	1.232	3.072	0.002
Wu et al. (2024)	1.271	0.278	0.077	0.726	1.815	4.575	0.000
Yilmaz and Karaoglan Yilmaz (2023)	0.920	0.309	0.096	0.314	1.526	2.977	0.003
	0.881	0.178	0.032	0.531	1.231	4.938	0.000



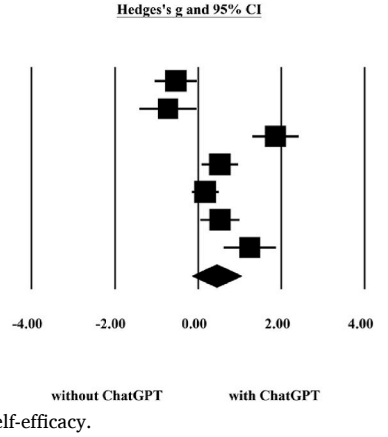
Appendix B2. Forest plots of effect sizes for affective-motivational states.

Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Chang et al. (2024)	1.409	0.312	0.097	0.797	2.020	4.515	0.000
Darmawansah et al. (2024)	0.613	0.247	0.061	0.128	1.097	2.477	0.013
Essel et al. (2024)	0.226	0.179	0.032	-0.124	0.576	1.266	0.206
Hu (2024b) - Comp.1	1.122	0.258	0.067	0.616	1.628	4.343	0.000
Hu (2024b) - Comp.2	2.236	0.309	0.096	1.630	2.842	7.230	0.000
Ji et al. (2023)	0.342	0.333	0.111	-0.311	0.994	1.026	0.305
Li (2023)	0.541	0.224	0.050	0.102	0.981	2.413	0.016
T. Li et al. (2024)	0.525	0.410	0.168	-0.278	1.328	1.282	0.200
Liao et al. (2024)	-0.007	0.296	0.088	-0.587	0.573	-0.023	0.981
Liu, Vobolevich and Oparin (2023)	1.636	0.113	0.013	1.416	1.857	14.533	0.000
Naamati-Schneider and Alt (2024)	0.117	0.163	0.027	-0.202	0.436	0.720	0.472
Shang and Geng (2024) - Comp.1	0.001	0.190	0.036	-0.372	0.374	0.007	0.995
Shang and Geng (2024) - Comp.2	0.418	0.192	0.037	0.041	0.795	2.175	0.030
Shang and Geng (2024) - Comp.3	0.540	0.194	0.038	0.161	0.920	2.789	0.005
Yilmaz and Karaoglan Yilmaz (2023)	0.921	0.309	0.096	0.315	1.527	2.979	0.003
	0.703	0.182	0.033	0.345	1.060	3.856	0.000

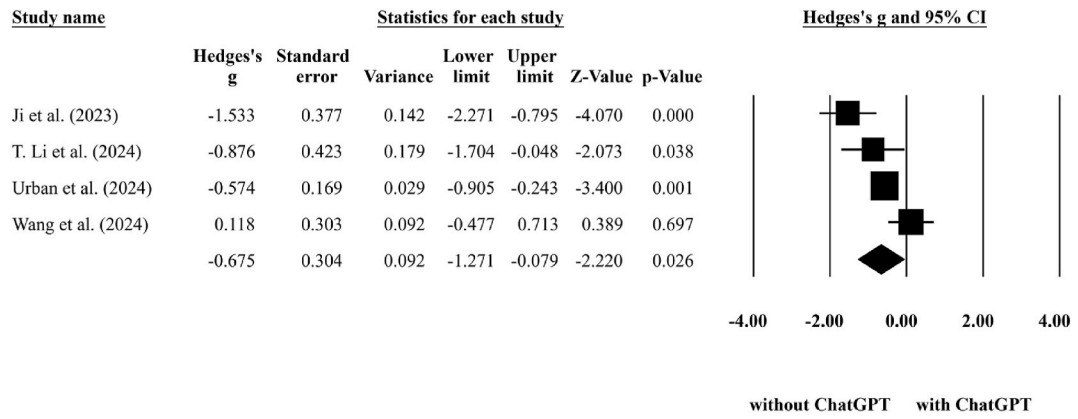


Appendix B3. Forest plots of effect sizes for higher-order thinking propensities.

Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Aydın Yıldız (2023)	-0.533	0.259	0.067	-1.041	-0.024	-2.053	0.040
Donald, Will and Christopher (2024)	-0.719	0.351	0.123	-1.407	-0.031	-2.047	0.041
Hu (2024a)	1.861	0.284	0.081	1.304	2.417	6.554	0.000
Li (2023)	0.526	0.224	0.050	0.087	0.966	2.349	0.019
Urban et al. (2024)	0.173	0.166	0.028	-0.152	0.498	1.042	0.297
Wu et al. (2023)	0.528	0.241	0.058	0.057	1.000	2.195	0.028
Yilmaz and Karaoglan Yilmaz (2023)	1.243	0.321	0.103	0.613	1.873	3.866	0.000
	0.441	0.297	0.088	-0.141	1.023	1.486	0.137



Appendix B4. Forest plots of effect sizes for self-efficacy.



Appendix B5. Forest plots of effect sizes for mental effort.

Data availability

Data will be made available on request.

References

Abraham, W. T., & Russell, D. W. (2008). Statistical power analysis in psychological research. *Social and Personality Psychology Compass, 2*(1), 283–301. https://doi.org/10.1111/j.1751-9004.2007.00052.x

Abu-Bader, S. H. (2021). *Using statistical methods in social science research* (3rd ed.). Oxford University Press.

Acosta-Enriquez, B. G., Arbulú Ballesteros, M. A., Huamaní Jordan, O., López Roca, C., & Saavedra Tirado, K. (2024). Analysis of college students' attitudes toward the use of ChatGPT in their academic activities: Effect of intent to use, verification of information and responsible use. *BMC Psychology, 12*(1), 255. https://doi.org/10.1186/s40359-024-01764-z

Adams, D., Chuah, K.-M., Devadason, E., & Azzis, M. S. A. (2023). From novice to navigator: Students' academic help-seeking behaviour, readiness, and perceived usefulness of ChatGPT in learning. *Education and Information Technologies, 29*, 13617–13634. https://doi.org/10.1007/s10639-023-12427-8

Adeshola, I., & Adepoju, A. P. (2023). *The opportunities and challenges of ChatGPT in education. Interactive Learning Environments*. Advance online publication. https://doi.org/10.1080/10494820.2023.2253858

* Ahmed Moneus, A. M., & Al-Wasy, B. Q. (2024). The impact of artificial intelligence on the quality of Saudi translators' performance. *Al-Andalus journal for Humanities & Social Sciences, 11*(96), 201–230. https://doi.org/10.35781/1637-000-096-006.

Al-khresheh, M. H. (2024). Bridging technology and pedagogy from a global lens: Teachers' perspectives on integrating ChatGPT in English language teaching. *Computers & Education: Artificial Intelligence, 6*, Article 100218. https://doi.org/10.1016/j.caeai.2024.100218

Al-Mamary, Y. H., Alfalah, A. A., Shamsuddin, A., & Abubakar, A. A. (2024). Artificial intelligence powering education: ChatGPT's impact on students' academic performance through the lens of technology-to-performance chain theory. *Journal of Applied Research in Higher Education, Advance online publication.* https://doi.org/10.1108/jarhe-04-2024-0179

Ali, D., Fatemi, Y., Boskabadi, E., Nikfar, M., Ugwuoke, J., & Ali, H. (2024). ChatGPT in teaching and learning: A systematic review. *Education Sciences, 14*(6), 643. https://doi.org/10.3390/educsci14060643

Almazrou, S., Alanezi, F., Almutairi, S. A., AboAlsamh, H. M., Alsedrah, I. T., Arif, W. M., Alsadhan, A. A., AlSanad, D. S., Alqahtani, N. S., AlShammary, M. H., Bakhshwain, A. M., Almuhanna, A. F., Almulhem, M., Alnaim, N., Albelali, S., & Attar, R. W. (2024). Enhancing medical students critical thinking skills through ChatGPT: An empirical study with medical students. *Nutrition and health*. Advance online publication. https://doi.org/10.1177/02601060241273627

* Almohesh, A. R. I. (2024). AI application (ChatGPT) and Saudi Arabian primary school students' autonomy in online classes: Exploring students and teachers' perceptions. *International Review of Research in Open and Distance Learning, 25*(3), 1–18. https://doi.org/10.19173/irrodl.v25i3.7641.

* Alneyadi, S., & Wardat, Y. (2023). ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools. *Contemporary Educational Technology, 15*(4), Article ep448. https://doi.org/10.30935/cedtech/13417.

* Alneyadi, S., & Wardat, Y. (2024). Integrating ChatGPT in grade 12 quantum theory education: An exploratory study at Emirate school (UAE). *International Journal of Information and Education Technology, 14*(3), 398–410. https://doi.org/10.18178/ijiet.2024.14.3.2061.

Amarathunga, B. (2024). ChatGPT in education: Unveiling frontiers and future directions through systematic literature review and bibliometric analysis. *Asian education and development studies*. Advance online publication. https://doi.org/10.1108/AEDS-05-2024-0101

* Ameen, L. T., Yousif, M. R., Alnoori, N. A. J., & Majeed, B. H. (2024). The impact of artificial intelligence on computational thinking in education at university. *International Journal of Engineering Pedagogy, 14*(5), 192–203. https://doi.org/10.3991/ijep.v14i5.49995.

Anderson, T., & Shattuck, J. (2012). Design-based research. *Educational Researcher, 41*(1), 16–25. https://doi.org/10.3102/0013189x11428813

Ansari, A. N., Ahmad, S., & Bhutta, S. M. (2024). Mapping the global evidence around the use of ChatGPT in higher education: A systematic scoping review. *Education and Information Technologies, 29*, 11281–11321. https://doi.org/10.1007/s10639-023-12223-4

* Avello-Martínez, R., Gajderowicz, T., & Gómez-Rodríguez, V. G. (2024). Is ChatGPT helpful for graduate students in acquiring knowledge about digital storytelling and reducing their cognitive load? An experiment. *Revista de Educación a Distancia, 24*(78), 8. https://doi.org/10.6018/red.604621.

* Aydın Yıldız, T. (2023). The impact of ChatGPT on language learners' motivation. *Journal of Teacher Education and Lifelong Learning, 5*(2), 582–597. https://doi.org/10.51535/tell.1314355.

Baig, M. I., & Yadegaridehkordi, E. (2024). ChatGPT in the higher education: A systematic literature review and research challenges. *International Journal of Educational Research, 127*, Article 102411. https://doi.org/10.1016/j.ijer.2024.102411

Baker, J. P., Goodboy, A. K., Bowman, N. D., & Wright, A. A. (2018). Does teaching with PowerPoint increase students' learning? A meta-analysis. *Computers & Education, 126*, 376–387. https://doi.org/10.1016/j.compedu.2018.08.003

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*(2), 122–147. https://doi.org/10.1037/0003-066X.37.2.122

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman and Company.

Barrett, A., & Pack, A. (2023). Not quite eye to A.I.: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education, 20*(1), 59. https://doi.org/10.1186/s41239-023-00427-0

* Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. *Humanities and Social Sciences Communications, 10*, 750. https://doi.org/10.1057/s41599-023-02269-7.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*(4), 1088–1101. https://doi.org/10.2307/2533446

* Beltozar-Clemente, S., & Díaz-Vega, E. (2024). Physics XP: Integration of ChatGPT and gamificatio to improve academic performance and motivation in physics 1 course. *International Journal of Engineering Pedagogy, 14*(6), 82–92. https://doi.org/10.3991/ijep.v14i6.47127.

Bhullar, P. S., Joshi, M., & Chugh, R. (2024). ChatGPT in higher education - a synthesis of the literature and a future research agenda. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12723-x

Biggs, J. B., & Tang, C. S. (2011). *Teaching for quality learning at university* (3rd ed.). Open University Press.

Borenstein, M. (2022). Comprehensive meta-analysis software. In M. Egger, J. P. T. Higgins, & G. D. Smith (Eds.), *Systematic reviews in health research: Meta-analysis in context* (3rd ed., pp. 535–548). John Wiley & Sons. https://doi.org/10.1002/9781119099369.ch27.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470743386

Boubker, O. (2024). From chatting to self-educating: Can AI tools boost student learning outcomes? *Expert Systems with Applications, 238*, Article 121820. https://doi.org/10.1016/j.eswa.2023.121820

* Boudouaia, A., Mouas, S., & Kouider, B. (2024). A study on ChatGPT-4 as an innovative approach to enhancing English as a foreign language writing learning. *Journal of Educational Computing Research, Advance online publication*. https://doi.org/10.1177/07356331241247465.

Bouteraa, M., Bin-Nashwan, S. A., Al-Daihani, M., Dirie, K. A., Benlahcene, A., Sadallah, M., Zaki, H. O., Lada, S., Ansar, R., Fook, L. M., & Chekima, B. (2024). Understanding the diffusion of AI-generative (ChatGPT) in higher education: Does students' integrity matter? *Computers in Human Behavior Reports, 14*, Article 100402. https://doi.org/10.1016/j.chbr.2024.100402

Bower, M., Torrington, J., Lai, J. W. M., Petocz, P., & Alfano, M. (2024). How should we change teaching and assessment in response to increasingly powerful generative artificial intelligence? Outcomes of the ChatGPT teacher survey. *Education and Information Technologies, 29*, 15403–15439. https://doi.org/10.1007/s10639-023-12405-0

Brom, C., Děchtěrenko, F., Frollová, N., Stárková, T., Bromová, E., & D'Mello, S. K. (2017). Enjoyment or involvement? Affective-Motivational mediation during learning from a complex computerized simulation. *Computers & Education, 114*, 236–254. https://doi.org/10.1016/j.compedu.2017.07.001

Bryman, A., & Bell, E. (2018). *Social research methods* (5th ed.). Oxford University Press.

Budhathoki, T., Zirar, A., Njoya, E. T., & Timsina, A. (2024). ChatGPT adoption and anxiety: A cross-country analysis utilising the unified theory of acceptance and use of technology (utaut). *Studies in Higher Education, 49*(5), 831–846. https://doi.org/10.1080/03075079.2024.2333937

Cambra-Fierro, J. J., Blasco, M. F., López-Pérez, M.-E. E., & Trifu, A. (2024). ChatGPT adoption and its influence on faculty well-being: An empirical research in higher education. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12871-0

Castillo-Manzano, J. I., Castro-Nuño, M., López-Valpuesta, L., Sanz-Díaz, M. T., & Yñiguez, R. (2016). Measuring the effect of ars on academic performance: A global meta-analysis. *Computers & Education, 96*, 109–121. https://doi.org/10.1016/j.compedu.2016.02.007

Çelik, F., Ersanlı, C. Y., & Arslanbay, G. (2024). Does AI simplification of authentic blog texts improve reading comprehension, inferencing, and anxiety? A one-shot intervention in Turkish efl context. *International Review of Research in Open and Distance Learning, 25*(3), 287–303. https://doi.org/10.19173/irrodl.v25i3.7779

Çelik, F., Yangın Ersanlı, C., & Arslanbay, G. (2024). Does AI simplification of authentic blog texts improve reading comprehension, inferencing, and anxiety? A one-shot intervention in Turkish efl context. *International Review of Research in Open and Distance Learning, 25*(3), 287–303. https://doi.org/10.19173/irrodl.v25i3.7779

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20*(1), 38. https://doi.org/10.1186/s41239-023-00408-3

Chan, C. K. Y., & Lee, K. K. W. (2023). The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X and millennial generation teachers? *Smart Learning Environments, 10*, 60. https://doi.org/10.1186/s40561-023-00269-3

Chan, C. K. Y., & Tsi, L. H. Y. (2024). Will generative AI replace teachers in higher education? A study of teacher and student perceptions. *Studies In Educational Evaluation, 83*. https://doi.org/10.1016/j.stueduc.2024.101395

* Chandrasekera, T., Hosseini, Z., & Perera, U. (2024). Can artificial intelligence support creativity in early design processes?. *International journal of architectural computing*. Advance online publication. https://doi.org/10.1177/14780771241254637.

* Chang, C.-Y., Yang, C.-L., Jen, H.-J., Ogata, H., & Hwang, G.-H. (2024). Facilitating nursing and health education by incorporating ChatGPT into learning designs. *Educational Technology & Society, 27*(1), 215–230. https://doi.org/10.30191/ETS.202401_27(1).TP02.

Chaudhry, I. S., Sarwary, S. A. M., El Refae, G. A., & Chabchoub, H. (2023). Time to revisit existing student's performance evaluation approach in higher education sector in a new era of ChatGPT — a case study. *Cogent Education, 10*(1), Article 2210461. https://doi.org/10.1080/2331186x.2023.2210461

* Chen, C.-H., & Chang, C.-L. (2024). Effectiveness of AI-assisted game-based learning on science learning outcomes, intrinsic motivation, cognitive load, and learning behavior. *Education and Information Technologies, 29*, 18621–18642. https://doi.org/10.1007/s10639-024-12553-x.

Chen, X., Hu, Z., & Wang, C. (2024). Empowering education development through aigc: A systematic literature review. *Education and Information Technologies, 29*, 17485–17537. https://doi.org/10.1007/s10639-024-12549-7

Chiarello, F., Giordano, V., Spada, I., Barandoni, S., & Fantoni, G. (2024). Future applications of generative large language models: A data-driven case study on ChatGPT. *Technovation, 133*, Article 103002. https://doi.org/10.1016/j.technovation.2024.103002

Chiasson, R. M., Goodboy, A. K., Vendemia, M. A., Beer, N., Meisz, G. C., Cooper, L., Arnold, A., Lincoski, A., George, W., Zuckerman, C., & Schrout, J. (2024). Does the human professor or artificial intelligence (AI) offer better explanations to students? Evidence from three within-subject experiments. In *Communication education*. Advance online publication. https://doi.org/10.1080/03634523.2024.2398105.

Chiu, T. K. F. (2023). The impact of generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and midjourney. In *Interactive learning environments*. Advance online publication. https://doi.org/10.1080/10494820.2023.2253861.

Cingillioglu, I., Gal, U., & Prokhorov, A. (2024). AI-experiments in education: An AI-driven randomized controlled trial for higher education research. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12633-y

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2024). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences, 121*(24), Article e2318124121. https://doi.org/10.1073/pnas.2318124121

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*(3), 444–452. https://doi.org/10.1007/s10956-023-10039-y

Coughlin, S. S. (1990). Recall bias in epidemiologic studies. *Journal of Clinical Epidemiology, 43*(1), 87–91. https://doi.org/10.1016/0895-4356(90)90060-3

Crawford, J., Allen, K.-A., Pani, B., & Cowling, M. (2024). When artificial intelligence substitutes humans in higher education: The cost of loneliness, student success, and retention. *Studies in Higher Education, 49*(5), 883–897. https://doi.org/10.1080/03075079.2024.2326956

Critical Appraisal Skills Programme. (2023). Randomised controlled trial. https://casp-uk.net/checklists/casp-rct-randomised-controlled-trial-checklist.pdf.

Croes, E. A. J., & Antheunis, M. L. (2021). Can we be friends with mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships, 38*(1), 279–300. https://doi.org/10.1177/0265407520959463

Dahri, N. A., Yahaya, N., & Al-Rahmi, W. M. (2024). Exploring the influence of ChatGPT on student academic success and career readiness. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-13148-2

Dai, Y., Lai, S., Lim, C. P., & Liu, A. (2023). ChatGPT and its impact on research supervision: Insights from Australian postgraduate research students. *Australasian Journal of Educational Technology, 39*(4), 74–88. https://doi.org/10.14742/ajet.8843

Dang, A., & Wang, H. (2024). Ethical use of generative AI for writing practices: Addressing linguistically diverse students in U.S. Universities' AI statements. *Journal of Second Language Writing, 66*, Article 101157. https://doi.org/10.1016/j.jslw.2024.101157
* Darmawansah, D., Rachman, D., Febiyani, F., & Hwang, G.-J. (2024). ChatGPT-supported collaborative argumentation: Integrating collaboration script and argument mapping to enhance EFL students' argumentation skills. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12986-4.
David, L., Biwer, F., Baars, M., Wijnia, L., Paas, F., & de Bruin, A. (2024). The relation between perceived mental effort, monitoring judgments, and learning outcomes: A meta-analysis. *Educational Psychology Review, 36*, 66. https://doi.org/10.1007/s10648-024-09903-z
de Winter, J. C. F. (2023). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education, Advance online publication*. https://doi.org/10.1007/s40593-023-00372-z
Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. *Frontiers in Education, 8*, 1–13. https://doi.org/10.3389/feduc.2023.1206936
Deng, R., Benckendorff, P., & Gannaway, D. (2019). Progress and new directions for teaching and learning in MOOCs. *Computers & Education, 129*, 48–60. https://doi.org/10.1016/j.compedu.2018.10.019
Deng, R., & Gao, Y. (2023). A review of eye tracking research on video-based learning. *Education and Information Technologies, 28*, 7671–7702. https://doi.org/10.1007/s10639-022-11486-7
Derakhshan, A., & Ghiasvand, F. (2024). Is ChatGPT an evil or an angel for second language education and research? A phenomenographic study of research-active efl teachers' perceptions. In *International journal of applied linguistics*. Advance online publication. https://doi.org/10.1111/ijal.12561.
Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences, 116*(39), 19251–19257. https://doi.org/10.1073/pnas.1821936116
Dianova, V. G., & Schultz, M. D. (2023). Discussing ChatGPT's implications for industry and higher education: The case for transdisciplinarity and digital humanities. *Industry and Higher Education, 37*(5), 593–600. https://doi.org/10.1177/09504222231199989
Dihan, Q., Chauhan, M. Z., Eleiwa, T. K., Brown, A. D., Hassan, A. K., Khodeiry, M. M., Elsheikh, R. H., Oke, I., Nihalani, B. R., VanderVeen, D. K., Sallam, A. B., & Elhusseiny, A. M. (2024). Large language models: A new frontier in paediatric cataract patient education. *British Journal of Ophthalmology, 108*(10), 1470–1476. https://doi.org/10.1136/bjo-2024-325252
Ding, L., Li, T., Jiang, S., & Gapud, A. (2023). Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education, 20*(1), 63. https://doi.org/10.1186/s41239-023-00434-1
* Donald, M. J., Will, D., & Christopher, M. E. (2024). Using ChatGPT with novice arduino programmers: Effects on performance, interest, self-efficacy, and programming ability. *Journal of Research in Technical Careers, 8*(1), 1–17. https://doi.org/10.9741/2578-2118.1152.
Dondio, P., Gusev, V., & Rocha, M. (2023). Do games reduce maths anxiety? A meta-analysis. *Computers & Education, 194*, Article 104650. https://doi.org/10.1016/j.compedu.2022.104650
Dube, S., Dube, S., Ndlovu, B. M., Maguraushe, K., Malungana, L., Kiwa, F. J., & Muduva, M. (2024). Students' perceptions of ChatGPT in education: A rapid systematic literature review. In K. Arai (Ed.), *Intelligent computing* (pp. 258–279). Springer. https://doi.org/10.1007/978-3-031-62273-1_18.
Dunnigan, J., Henriksen, D., Mishra, P., & Lake, R. (2023). "Can we just please slow it all down?" School leaders take on ChatGPT. *TechTrends, 67*(6), 878–884. https://doi.org/10.1007/s11528-023-00914-1
Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235
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
Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
ElSayary, A. (2023). An investigation of teachers' perceptions of using ChatGPT as a supporting tool for teaching and learning in the digital era. *Journal of Computer Assisted Learning, 40*(3), 931–945. https://doi.org/10.1111/jcal.12926
* Emran, A. Q., Mohammed, M. N., Saeed, H., Abu Keir, M. Y., Alani, Z. N., & Mohammed Ibrahim, F. (2024). *Paraphrasing ChatGPT answers as a tool to enhance university students' academic writing skills 2024*. Bahrain: ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems.
* Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers & Education: Artificial Intelligence, 6*, Article 100198. https://doi.org/10.1016/j.caeai.2023.100198.
* Essien, A., Bukoye, O. T., O'Dea, X., & Kremantzis, M. (2024). The influence of AI text generators on critical thinking skills in UK business schools. *Studies in Higher Education, 49*(5), 865–882. https://doi.org/10.1080/03075079.2024.2316881.
* Farah, J. C., Ingram, S., Spaenlehauer, B., Lasne, F. K.-L., & Gillet, D. (2023). *Prompting large language models to power educational chatbots International Conference on Web-Based Learning*. Sydney.
Flodén, J. (2024). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. In *British educational research journal*. Advance online publication https://doi.org/10.1002/berj.4069.
Galindo-Domínguez, H., Delgado, N., Losada, D., & Etxabe, J.-M. (2023). An analysis of the use of artificial intelligence in education in Spain: The in-service teacher's perspective. *Journal of Digital Learning in Teacher Education, 40*(1), 41–56. https://doi.org/10.1080/21532974.2023.2284726
Gammoh, L. A. (2024). ChatGPT risks in academia: Examining university educators' challenges in Jordan. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-13009-y
* Gan, W., Ouyang, J., Li, H., Xue, Z., Zhang, Y., Dong, Q., Huang, J., Zheng, X., & Zhang, Y.. Integrating ChatGPT in orthopedic education for medical undergraduates: Randomized controlled trial. *Journal of Medical Internet Research, 26*, Article e57037. https://doi.org/10.2196/57037.
* Gao, S. (2024). Can artificial intelligence give a hand to open and distributed learning? A probe into the state of undergraduate students' academic emotions and test anxiety in learning via ChatGPT. *International Review of Research in Open and Distance Learning, 25*(3), 199–218. https://doi.org/10.19173/irrodl.v25i3.7742.
Gao, L., López-Pérez, M. E., Melero-Polo, I., & Trifu, A. (2024). Ask ChatGPT first! Transforming learning experiences in the age of artificial intelligence. *Studies in Higher Education, Advance online publication*. https://doi.org/10.1080/03075079.2024.2323571
Gao, Q., Yan, Z., Zhao, C., Pan, Y., & Mo, L. (2014). To ban or not to ban: Differences in mobile phone policies at elementary, middle, and high schools. *Computers in Human Behavior, 38*, 25–32. https://doi.org/10.1016/j.chb.2014.05.011
Garcia Castro, R. A., Mayta Cachicatari, N. A., Bartesaghi Aste, W. M., & Llapa Medina, M. P. (2024). Exploration of ChatGPT in basic education: Advantages, disadvantages, and its impact on school tasks. *Contemporary Educational Technology, 16*(3), ep511. https://doi.org/10.30935/cedtech/14615
García, V. A., la Fuente, S. D.-d., Martín, J. I. S., & Ordax, J. M. G. (2024). *A critical approach to the use of ChatGPT in higher education*. Cham: The 17th International Conference on Industrial Engineering and Industrial Management.
Garofalo, S. G., & Farenga, S. J. (2024). Science teacher perceptions of the state of knowledge and education at the advent of generative artificial intelligence popularity. *Science & education*. Advance online publication. https://doi.org/10.1007/s11191-024-00534-y
Gencer, G., & Gencer, K. (2024). A comparative analysis of ChatGPT and medical faculty graduates in medical specialization exams: Uncovering the potential of artificial intelligence in medical education. *Cureus, 16*(8), Article e66517. https://doi.org/10.7759/cureus.66517
Gorard, S., & Cook, T. (2007). Where does good evidence come from? *International Journal of Research and Method in Education, 30*(3), 307–323. https://doi.org/10.1080/17437270701614790

Grájeda, A., Córdova, P., Córdova, J. P., Laguna-Tapia, A., Burgos, J., Rodríguez, L., Arandia, M., & Sanjinés, A. (2024). Embracing artificial intelligence in the arts classroom: Understanding student perceptions and emotional reactions to AI tools. *Cogent Education, 11*(1), Article 2378271. https://doi.org/10.1080/2331186x.2024.2378271

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences, 13*(7). https://doi.org/10.3390/educsci13070692

Grassini, S., Aasen, M. L., & Møgelvang, A. (2024). Understanding university students' acceptance of ChatGPT: Insights from the UTAUT2 model. *Applied Artificial Intelligence, 38*(1), Article 2371168. https://doi.org/10.1080/08839514.2024.2371168

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*(1), 1–21. https://doi.org/10.1080/00401706.1969.10490657

Haig, B. D. (2003). What is a spurious correlation? *Understanding Statistics, 2*(2), 125–132. https://doi.org/10.1207/S15328031US0202_03

Haindl, P., & Weinberger, G. (2024). Students' experiences of using ChatGPT in an undergraduate programming course. *IEEE Access, 12*, 43519–43529. https://doi.org/10.1109/access.2024.3380909

Halme, P., Toivanen, T., Honkanen, M., Kotiaho, J. S., MönkkÖnen, M., & Timonen, J. (2010). Flawed meta-analysis of biodiversity effects of forest management. *Conservation Biology, 24*(4), 1154–1156. https://doi.org/10.1111/j.1523-1739.2010.01542.x

Hamerman, E. J., Aggarwal, A., & Martins, C. (2024). An investigation of generative AI in the classroom and its implications for university policy. *Quality assurance in education*. Advance online publication. https://doi.org/10.1108/qae-08-2024-0149

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332

Hays, L., Jurkowski, O., & Sims, S. K. (2024). ChatGPT in K-12 education. *TechTrends, 68*, 281–294. https://doi.org/10.1007/s11528-023-00924-z

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

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

Hehir, E., Zeller, M., Luckhurst, J., & Chandler, T. (2021). Developing student connectedness under remote learning using digital resources: A systematic review. *Education and Information Technologies, 26*(5), 6531–6548. https://doi.org/10.1007/s10639-021-10577-1

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports, 13*, Article 18617. https://doi.org/10.1038/s41598-023-45644-9

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). John Wiley & Sons.

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

Higgs, J. M., & Stornaiuolo, A. (2024). Being human in the age of generative AI: Young people's ethical concerns about writing and living with machines. *Reading Research Quarterly, 59*(4), 632–650. https://doi.org/10.1002/rrq.552

Howitt, D., & Cramer, D. (2017). *Understanding statistics in psychology with SPSS* (7th ed.). Pearson.

* Hsu, M.-H. (2023). Mastering medical terminology with ChatGPT and Termbot. *Health Education Journal, Advance online publication*. https://doi.org/10.1177/00178969231197371.

* Hu, Y.-H. (2024a). Implementing generative AI chatbots as a decision aid for enhanced values clarification exercises in online business ethics education. *Educational Technology & Society, 27*(3), 356–373. https://doi.org/10.30191/ETS.202407_27(3).TP02.

* Hu, Y.-H. (2024b). Improving ethical dilemma learning: Featuring thinking aloud pair problem solving (TAPPS) and AI-assisted virtual learning companion. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12754-4.

* Huesca, G., Martínez-Treviño, Y., Molina-Espinosa, J. M., Sanromán-Calleros, A. R., Martínez-Román, R., Cendejas-Castro, E. A., & Bustos, R. (2024). Effectiveness of using ChatGPT as a tool to strengthen benefits of the flipped learning strategy. *Education Sciences, 14*(6), 660. https://doi.org/10.3390/educsci14060660.

Hyun Baek, T., & Kim, M. (2023). Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. *Telematics and Informatics, 83*, Article 102030. https://doi.org/10.1016/j.tele.2023.102030

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology, 15*(4), ep464. https://doi.org/10.30935/cedtech/13605

İpek, Z. H., Gözüm, A.İ. C., Papadakis, S., & Kallogiannakis, M. (2023). Educational applications of the ChatGPT AI system: A systematic review research. *Educational Process: International Journal, 12*(3), 26–55. https://doi.org/10.22521/edupij.2023.123.2

* Ironsi, C. S., & Ironsi, S. S. (2024). Experimental evidence for the efficacy of generative AI in improving students' writing skills. *Quality assurance in education*. Advance online publication. https://doi.org/10.1108/qae-04-2024-0065.

Islam, I., & Islam, M. N. (2024). Exploring the opportunities and challenges of ChatGPT in academia. *Discover Education, 3*(1), 31. https://doi.org/10.1007/s44217-024-00114-w

Jaboob, M., Hazaimeh, M., & Al-Ansi, A. M. (2024). Integration of generative AI techniques and applications in student behavior and cognitive achievement in Arab higher education. *International Journal of Human–Computer Interaction, Advance online punlication*. https://doi.org/10.1080/10447318.2023.2300016

Jahic, I., Ebner, M., & Schön, S. (2023). *Harnessing the power of artificial intelligence and ChatGPT in education – a first rapid literature review*. Vienna, Austria: EdMedia + Innovate Learning. https://www.learntechlib.org/p/222670/.

Jarry Trujillo, C., Vela Ulloa, J., Escalona Vivas, G., Grasset Escobar, E., Villagrán Gutiérrez, I., Achurra Tirado, P., & Varas Cohen, J. (2024). Surgeons vs ChatGPT: Assessment and feedback performance based on real surgical scenarios. *Journal of Surgical Education, 81*(7), 960–966. https://doi.org/10.1016/j.jsurg.2024.03.012

Jensen, L. X., Buhl, A., Sharma, A., & Bearman, M. (2024). Generative AI and higher education: A review of claims from the first months of ChatGPT. *Higher education*. Advance online publication. https://doi.org/10.1007/s10734-024-01265-3

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 28*(12), 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

* Ji, Y., Zou, X., Li, T., & Zhan, Z. (2023). *The effectiveness of ChatGPT on pre-service teachers' STEM teaching literacy, learning performance, and cognitive load in a teacher training course*. Guangzhou, China: The 6th International Conference on Educational Technology Management.

Jiang, Y., Xie, L., Lin, G., & Mo, F. (2024). *Widen the debate: What is the academic community's perception on ChatGPT? Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12677-0

Jo, H. (2023). Understanding AI tool engagement: A study of ChatGPT usage and word-of-mouth among university students and office workers. *Telematics and Informatics, 85*, Article 102067. https://doi.org/10.1016/j.tele.2023.102067

Jo, H. (2024). From concerns to benefits: A comprehensive study of ChatGPT usage in education. *International Journal of Educational Technology in Higher Education, 21*(1), 35. https://doi.org/10.1186/s41239-024-00471-4

Jochim, J., & Lenz-Kesekamp, V. K. (2024). Teaching and testing in the era of text-generative AI: Exploring the needs of students and teachers. *Information and learning sciences*. Advance online publication. https://doi.org/10.1108/ils-10-2023-0165

Jošt, G., Taneski, V., & Karakatič, S. (2024). The impact of large language models on programming education and student learning outcomes. *Applied Sciences, 14*(10). https://doi.org/10.3390/app14104115

Kapur, M., Saba, J., & Roll, I. (2023). Prior math achievement and inventive production predict learning from productive failure. *Npj Science of Learning, 8*(1), 15. https://doi.org/10.1038/s41539-023-00165-y

Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12574-6

Karkoulian, S., Sayegh, N., & Sayegh, N. (2024). ChatGPT unveiled: Understanding perceptions of academic integrity in higher education - a qualitative approach. *Journal of Academic Ethics, Advance online publication*. https://doi.org/10.1007/s10805-024-09543-6

* Kavadella, A., Dias da Silva, M. A., Kaklamanos, E. G., Stamatopoulos, V., & Giannakopoulos, K. (2024). Evaluation of ChatGPT's real-life implementation in undergraduate dental education: Mixed methods study. *JMIR Medical Education, 10*, Article e51344. https://doi.org/10.2196/51344.

* Khuibut, W., Premthaisong, S., & Chaipidech, P. (2023). Integrating ChatGPT into synectics model to improve high school student's creative writing skill. *The 31st International Conference on Computers in Education, Matsue, Japan*.
* Kim, R. (2024). Effects of ChatGPT on Korean EFL learners' main-idea reading comprehension via top-down processing. *Language Research, 60*(1), 83–106. https://doi.org/10.30961/lr.2024.60.1.83.
Kim, J., Ham, Y., & Lee, S.-S. (2024). Differences in student-AI interaction process on a drawing task: Focusing on students' attitude towards AI and the level of drawing skills. *Australasian Journal of Educational Technology, 40*(1), 19–41. https://doi.org/10.14742/ajet.8859
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.
Koltovskaia, S., Rahmati, P., & Saeli, H. (2024). Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement. *Journal of Second Language Writing, 65*, Article 101130. https://doi.org/10.1016/j.jslw.2024.101130
Korseberg, L., & Elken, M. (2024). Waiting for the revolution: How higher education institutions initially responded to ChatGPT. *Higher education*. Advance online publication. https://doi.org/10.1007/s10734-024-01256-4
* Kosar, T., Ostojić, D., Liu, Y. D., & Mernik, M. (2024). Computer science education in ChatGPT era: Experiences from an experiment in a programming course for novice programmers. *Mathematics, 12*(5), 629. https://doi.org/10.3390/math12050629.
Krieglstein, F., Beege, M., Rey, G. D., Ginns, P., Krell, M., & Schneider, S. (2022). A systematic meta-analysis of the reliability and validity of subjective cognitive load questionnaires in experimental multimedia learning research. *Educational Psychology Review, 34*(4), 2485–2541. https://doi.org/10.1007/s10648-022-09683-4
* Kucuk, T. (2024). Chatgpt integrated grammar teaching and learning in efl classes: A study on tishk international university students in erbil, Iraq. *Arab World English Journal*, 100–111. https://doi.org/10.24093/awej/ChatGPT.6.
Kwan, V. S. Y., John, O. P., Robins, R. W., & Kuang, L. L. (2008). Conceptualizing and assessing self-enhancement bias: A componential approach. *Journal of Personality and Social Psychology, 94*(6), 1062–1077. https://doi.org/10.1037/0022-3514.94.6.1062
Laak, K.-J., & Aru, J. (2024). *Generative AI in K-12: Opportunities for learning and utility for teachers 25th international conference, AIED*. Brazil: Recife.
Al-Qaysi, N., Al-Emran, M., Al-Sharafi, M. A., Iranmanesh, M., Ahmad, A., & Mahmoud, M. A. (2024). Determinants of ChatGPT use and its impact on learning performance: An integrated model of BRT and TPB. *International Journal of Human–Computer Interaction*, Advance online publication. https://doi.org/10.1080/10447318.2024.2361210.
Lambert, J., & Stevens, M. ChatGPT and generative AI technology: A mixed bag of concerns and new opportunities. Computers in the Schools, Advance online publication. https://doi.org/10.1080/07380569.2023.2256710.
Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open, 6*, Article 100174. https://doi.org/10.1016/j.caeo.2024.100174
* Lee, H.-Y., Chen, P.-H., Wang, W.-S., Huang, Y.-M., & Wu, T.-T. (2024). Empowering ChatGPT with guidance mechanism in blended learning: Effect of self-regulated learning, higher-order thinking skills, and knowledge construction. *International Journal of Educational Technology in Higher Education, 21*(1), 16. https://doi.org/10.1186/s41239-024-00447-4.
Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies, 29*, 11483–11515. https://doi.org/10.1007/s10639-023-12249-8
Lee, M., Oi-yeung Lam, B., Ju, E., & Dean, J. (2017). Part-time employment and problem behaviors: Evidence from adolescents in South Korea. *Journal of Research on Adolescence, 27*(1), 88–104. https://doi.org/10.1111/jora.12258
Lee, L., Packer, T. L., Tang, S. H., & Girdler, S. (2008). Self-management education programs for age-related macular degeneration: A systematic review. *Australasian Journal on Ageing, 27*(4), 170–176. https://doi.org/10.1111/j.1741-6612.2008.00298.x
Lee, S., & Song, K.-S. (2024). Teachers' and students' perceptions of AI-generated concept explanations: Implications for integrating generative AI in computer science education. *Computers & Education: Artificial Intelligence, 7*, Article 100283. https://doi.org/10.1016/j.caeai.2024.100283
Lee, G.-G., & Zhai, X. (2024). Using ChatGPT for science learning: A study on pre-service teachers' lesson planning. *IEEE Transactions on Learning Technologies, 17*, 1683–1700. https://doi.org/10.1109/tlt.2024.3401457
* Li, H. (2023). Effects of a ChatGPT-based flipped learning guiding approach on learners' courseware project performances and perceptions. *Australasian Journal of Educational Technology, 39*(5), 40–58. https://doi.org/10.14742/ajet.8923.
* Li, T., Ji, Y., & Zhan, Z. (2024). Expert or machine? Comparing the effect of pairing student teacher with in-service teacher and ChatGPT on their critical thinking, learning performance, and cognitive load in an integrated-STEM course. *Asia Pacific Journal of Education, 44*(1), 45–60. https://doi.org/10.1080/02188791.2024.2305163.
Li, L., Ma, Z., Fan, L., Lee, S., Yu, H., & Hemphill, L. (2024). ChatGPT in education: A discourse analysis of worries and concerns on social media. *Education and Information Technologies, 29*, 10729–10762. https://doi.org/10.1007/s10639-023-12256-9
* Li, Y., Sadiq, G., Qambar, G., & Zheng, P. (2024). The impact of students' use of ChatGPT on their research skills: The mediating effects of autonomous motivation, engagement, and self-directed learning. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12981-9.
Lian, Y., Tang, H., Xiang, M., & Dong, X. (2024). Public attitudes and sentiments toward ChatGPT in China: A text mining analysis based on social media. *Technology in Society, 76*, Article 102442. https://doi.org/10.1016/j.techsoc.2023.102442
* Liao, J., Zhong, L., Zhe, L., Xu, H., Liu, M., & Xie, T. (2024). Scaffolding computational thinking with ChatGPT. *IEEE Transactions on Learning Technologies, 17*, 1668–1682. https://doi.org/10.1109/tlt.2024.3392896.
Lin, Z., & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System, 123*, Article 103344. https://doi.org/10.1016/j.system.2024.103344
Liu, X. (2024). Navigating uncharted waters: Teachers' perceptions of and reactions to AI-induced challenges to assessment. *The asia-pacific education researcher*. Advance online publication. https://doi.org/10.1007/s40299-024-00890-x
* Liu, Z., Vobolevich, A., & Oparin, A. (2023). The influence of AI ChatGPT on improving teachers' creative thinking. *International Journal of Learning, Teaching and Educational Research, 22*(12), 124–139. https://doi.org/10.26803/ijlter.22.12.7.
Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education, 211*. https://doi.org/10.1016/j.compedu.2023.104977
Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4). https://doi.org/10.3390/educsci13040410
Lo, C. K., Hew, K. F., & Jong, M. S.-y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education, 105100*. https://doi.org/10.1016/j.compedu.2024.105100
* Lu, J., Zheng, R., Gong, Z., & Xu, H. (2024). Supporting teachers' professional development with generative AI: The effects on higher order thinking and self-efficacy. *IEEE Transactions on Learning Technologies, 17*, 1279–1289. https://doi.org/10.1109/tlt.2024.3369690.
* Lyu, W., Wang, Y., Chung, T. R., Sun, Y., & Zhang, Y. (2024). *Evaluating the effectiveness of llMs in introductory computer science education: A semester-long field study*. Atlanta: The 11th ACM Conference on Learning @ Scale.
Ma, Q., Crosthwaite, P., Sun, D., & Zou, D. (2024). Exploring ChatGPT literacy in language education: A global perspective and comprehensive approach. *Computers & Education: Artificial Intelligence, 7*, Article 100278. https://doi.org/10.1016/j.caeai.2024.100278
Maatouk, A., Piovesan, N., Ayed, F., Domenico, A. D., & Debbah, M. (2024). Large language models for telecom: Forthcoming impact on the industry. In *IEEE communications magazine*. Advance online publication. https://doi.org/10.1109/mcom.001.2300473.
* Maghamil, M. C., & Sieras, S. G. (2024). Impact of ChatGPT on the academic writing quality of senior high school students. *Journal of English Language Teaching & Applied Linguistics, 6*(2), 115–128. https://doi.org/10.32996/jeltal.2024.6.2.14.
* Mahapatra, S. (2024). Impact of ChatGPT on esl students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*(1), 9. https://doi.org/10.1186/s40561-024-00295-9.
Maheshwari, G. (2024). Factors influencing students' intention to adopt and use ChatGPT in higher education: A study in the Vietnamese context. *Education and Information Technologies, 29*(10), 12167–12195. https://doi.org/10.1007/s10639-023-12333-z

Mahrishi, M., Abbas, A., Radovanovic, D., & Hosseini, S. (2024). Emerging dynamics of ChatGPT in academia: A scoping review. *Journal of University Teaching and Learning Practice, 21*(1–31).
Mai, D. T. T., Da, C. V., & Hanh, N. V. (2024). The use of ChatGPT in teaching and learning: A systematic review through swot analysis approach. *Frontiers in Education, 9*, Article 1328769. https://doi.org/10.3389/feduc.2024.1328769
McGrath, C., Farazouli, A., & Cerratto-Pargman, T. (2024). Generative AI chatbots in higher education: A review of an emerging research area. *Higher education*. Advance online publication. https://doi.org/10.1007/s10734-024-01288-w
McHugh, M. L. (2012). Interrater reliability: The Kappa statistic. *Biochemia Medica, 22*(3), 276–282.
* Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers & Education: Artificial Intelligence, 6*, Article 100199. https://doi.org/10.1016/j.caeai.2023.100199.
Mohamed, A. M. (2024). Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a foreign language (EFL) teaching: Perceptions of EFL faculty members. *Education and Information Technologies, 29*(3), 3195–3217. https://doi.org/10.1007/s10639-023-11917-z
Monib, W. K., Qazi, A., & Mahmud, M. M. (2024). Exploring learners' experiences and perceptions of ChatGPT as a learning tool in higher education. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-13065-4
Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J. M., & López-Meneses, E. (2023). Impact of the implementation of ChatGPT in education: A systematic review. *Computers, 12*(8), 153. https://doi.org/10.3390/computers12080153
Moorhouse, B. L., & Kohnke, L. (2024). The effects of generative AI on initial language teacher education: The perceptions of teacher educators. *System, 122*, Article 103290. https://doi.org/10.1016/j.system.2024.103290
Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open, 5*. https://doi.org/10.1016/j.caeo.2023.100151
* Mugableh, A. I. (2024). The impact of ChatGPT on the development of vocabulary knowledge of Saudi EFL students. *Arab World English Journal*, 265–281. https://doi.org/10.24093/awej/ChatGPT.18.
* Mun, C.-y. (2024). EFL learners' English writing feedback and their perception of using ChatGPT. *Journal of English Teaching through Movies and Media, 25*(2), 26–39.
Mutlu-Bayraktar, D., Cosgun, V., & Altan, T. (2019). Cognitive load in multimedia learning environments: A systematic review. *Computers & Education, 141*. https://doi.org/10.1016/j.compedu.2019.103618
Na, H., Staudt Willet, K. B., Shi, H., Hur, J., He, D., & Kim, C. (2024). Initial discussions of ChatGPT in education-related subreddits. *Journal of Research on Technology in Education, Advance online publication*. https://doi.org/10.1080/15391523.2024.2338091
* Naamati-Schneider, L., & Alt, D. (2024). Beyond digital literacy: The era of AI-powered assistants and evolving user skills. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12694-z.
Nam, B. H., & Bai, Q. (2023). ChatGPT and its ethical implications for stem research and higher education: A media discourse analysis. *International Journal of STEM Education, 10*(1), 66. https://doi.org/10.1186/s40594-023-00452-5
Newton, P., & Xiromeriti, M. (2024). ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. *Assessment & Evaluation in Higher Education, 49*(6), 781–798. https://doi.org/10.1080/02602938.2023.2299059
* Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology, 55*(4), 1328–1353. https://doi.org/10.1111/bjet.13454.
Ngo, T. T. A., Tran, T. T., An, G. K., & Nguyen, P. T. (2024). ChatGPT for educational purposes: Investigating the impact of knowledge management factors on student satisfaction and continuous usage. *IEEE Transactions on Learning Technologies, 17*, 1367–1378. https://doi.org/10.1109/tlt.2024.3383773
NHLBI. (2021). Study quality assessment tools: Quality assessment of controlled intervention studies. https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools.
* Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2023). Is chatgpt a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning, 40*(2), 919–930. https://doi.org/10.1111/jcal.12929.
Niloy, A. C., Bari, M. A., Sultana, J., Chowdhury, R., Raisa, F. M., Islam, A., Mahmud, S., Jahan, I., Sarkar, M., Akter, S., Nishat, N., Afroz, M., Sen, A., Islam, T., Tareq, M. H., & Hossen, M. A. (2024). Why do students use ChatGPT? Answering through a triangulation approach. *Computers & Education: Artificial Intelligence, 6*, Article 100208. https://doi.org/10.1016/j.caeai.2024.100208
Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586
Ofem, U. J., Owan, V. J., Iyam, M. A., Udeh, M. I., Anake, P. M., & Ovat, S. V. (2024). Students' perceptions, attitudes and utilisation of ChatGPT for academic dishonesty: Multigroup analyses via PLS−SEM. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12850-5
Okulu, H. Z., & Muslu, N. (2024). Designing a course for pre-service science teachers using ChatGPT: What ChatGPT brings to the table. *Interactive learning environments*. Advance online publication. https://doi.org/10.1080/10494820.2024.2322462
Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2010). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63–71. https://doi.org/10.1207/s15326985ep3801_8
Paas, F., Tuovinen, J. E., van Merriënboer, J. J. G., & Aubteen Darabi, A. (2005). A motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction. *Educational Technology Research & Development, 53*(3), 25–34. https://doi.org/10.1007/BF02504795
Pack, A., & Maloney, J. (2023). Using generative artificial intelligence for language education research: Insights from using OpenAI's ChatGPT. *Tesol Quarterly, 57*(4), 1571–1582. https://doi.org/10.1002/tesq.3253
Pack, A., & Maloney, J. (2024). Using artificial intelligence in TESOL: Some ethical and pedagogical considerations. *Tesol Quarterly, 58*(2), 1007–1018. https://doi.org/10.1002/tesq.3320
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. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71
Park, Y., & Doo, M. Y. (2024). Role of AI in blended learning: A systematic literature review. *International Review of Research in Open and Distance Learning, 25*(1), 164–196. https://doi.org/10.19173/irrodl.v25i1.7566
Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X.
Peng, C.-Y. J., Long, H., & Abaci, S. (2012). Power analysis software for educational researchers. *The Journal of Experimental Education, 80*(2), 113–136. https://doi.org/10.1080/00220973.2011.647115
Perera, P., & Lankathilaka, M. (2023). AI in higher education: A literature review of ChatGPT and guidelines for responsible implementation. *International Journal of Research and Innovation in Social Science, 7*(6), 306–314. https://doi.org/10.47772/IJRISS.2023.7623
Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2023). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics, 22*(1), 89–113. https://doi.org/10.1007/s10805-023-09492-6
Playfoot, D., Quigley, M., & Thomas, A. G. (2024). Hey ChatGPT, give me a title for a paper about degree apathy and student use of AI for assignment writing. *The Internet and Higher Education, 62*, Article 100950. https://doi.org/10.1016/j.iheduc.2024.100950
Polat, H., Topuz, A. C., Yıldız, M., Taşlıbeyaz, E., & Kurşun, E. (2024). A bibliometric analysis of research on ChatGPT in education. *International Journal of Technology in Education, 7*(1), 59–85. https://doi.org/10.46328/ijte.606
Polyportis, A. (2024). A longitudinal study on artificial intelligence adoption: Understanding the drivers of ChatGPT usage behavior change in higher education. *Frontiers in Artificial Intelligence, 6*, Article 1324398. https://doi.org/10.3389/frai.2023.1324398
Pradana, M., Elisa, H. P., & Syarifuddin, S. (2023). Discussing ChatGPT in education: A literature review and bibliometric analysis. *Cogent Education, 10*(2), Article 2243134. https://doi.org/10.1080/2331186x.2023.2243134

Rawas, S. (2024). ChatGPT: Empowering lifelong learning in the digital age of higher education. *Education and Information Technologies, 29*(6), 6895–6908. https://doi.org/10.1007/s10639-023-12114-8

Rienties, B., Domingue, J., Duttaroy, S., Herodotou, C., Tessarolo, F., & Whitelock, D. (2024). What distance learning students want from an AI Digital Assistant. *Distance education*. Advance online publication. https://doi.org/10.1080/01587919.2024.2338717

Roohr, K., Olivera-Aguilar, M., Ling, G., & Rikoon, S. (2019). A multi-level modeling approach to investigating students' critical thinking at higher education institutions. *Assessment & Evaluation in Higher Education, 44*(6), 946–960. https://doi.org/10.1080/02602938.2018.1556776

Salifu, I., Arthur, F., Arkorful, V., Abam Nortey, S., & Solomon Osei-Yaw, R. (2024). Economics students' behavioural intention and usage of ChatGPT in higher education: A hybrid structural equation modelling-artificial neural network approach. *Cogent Social Sciences, 10*(1), Article 2300177. https://doi.org/10.1080/23311886.2023.2300177

Sallam, M., Al-Mahzoum, K., Almutawaa, R. A., Alhashash, J. A., Dashti, R. A., AlSafy, D. R., Almutairi, R. A., & Barakat, M. (2024). The performance of OpenAI ChatGPT-4 and google gemini in virology multiple-choice questions: A comparative analysis of English and Arabic responses. *BMC Research Notes, 17*(1), 247. https://doi.org/10.1186/s13104-024-06920-7

Samala, A. D., Rawas, S., Wang, T., Reed, J. M., Kim, J., Howard, N.-J., & Ertz, M. (2024). Unveiling the landscape of generative artificial intelligence in education: A comprehensive taxonomy of applications, challenges, and future prospects. *Education and Information Technologies, Advance online publication.*. https://doi.org/10.1007/s10639-024-12936-0

Sandmann, S., Riepenhausen, S., Plagwitz, L., & Varghese, J. (2024). Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nature Communications, 15*(1), 2050. https://doi.org/10.1038/s41467-024-46411-8

Saritepeci, M., & Yildiz Durak, H. (2024). Effectiveness of artificial intelligence integration in design-based learning on design thinking mindset, creative and reflective thinking skills: An experimental study. *Education and information technologies*. Advance Online Publication. https://doi.org/10.1007/s10639-024-12829-2

Schmidt, S. A. J., Lo, S., & Hollestein, L. M. (2018). Research techniques made simple: Sample size estimation and power calculation. *Journal of Investigative Dermatology, 138*(8), 1678–1682. https://doi.org/10.1016/j.jid.2018.06.165

Schott, C., van Roekel, H., & Tummers, L. G. (2020). Teacher leadership: A systematic review, methodological quality assessment and conceptual framework. *Educational Research Review, 31*, Article 100352. https://doi.org/10.1016/j.edurev.2020.100352

Schroeder, N. L., Siegle, R. F., & Craig, S. D. (2023). A meta-analysis on learning from 360° video. *Computers & Education, 206*, Article 104901. https://doi.org/10.1016/j.compedu.2023.104901

Šedlbauer, J., Činčera, J., Slavík, M., & Hartlová, A. (2024). Students' reflections on their experience with ChatGPT. *Journal of Computer Assisted Learning, 40*, 1526–1534. https://doi.org/10.1111/jcal.12967

Seth, I., Lim, B., Cevik, J., Sofiadellis, F., Ross, R. J., Cuomo, R., & Rozen, W. M. (2024). Utilizing GPT-4 and generative artificial intelligence platforms for surgical education: An experimental study on skin ulcers. *European Journal of Plastic Surgery, 47*(1), 19. https://doi.org/10.1007/s00238-024-02162-9

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Shahzad, M. F., Xu, S., & Zahid, H. (2024). Exploring the impact of generative AI-based technologies on learning performance through self-efficacy, fairness & ethics, creativity, and trust in higher education. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12949-9

* Shang, S., & Geng, S. (2024). Empowering learners with AI-generated content for programming learning and computational thinking: The lens of extended effective use theory. *Journal of Computer Assisted Learning, 40*, 1941–1958. https://doi.org/10.1111/jcal.12996.

* Shi, S. J., Li, J. W., & Zhang, R. (2024). A study on the impact of Generative Artificial Intelligence supported Situational Interactive Teaching on students' 'flow' experience and learning effectiveness — a case study of legal education in China. *Asia Pacific Journal of Education, 44*(1), 112–138. https://doi.org/10.1080/02188791.2024.2305161.

* Shin, H., De Gagne, J. C., Kim, S. S., & Hong, M. (2024). The impact of artificial intelligence-assisted learning on nursing students' ethical decision-making and clinical reasoning in pediatric care: A quasi-experimental study. *CIN: Computers, Informatics, Nursing, 42*(10), 704–711. https://doi.org/10.1097/cin.0000000000001177.

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12817-6

* Silitonga, L. M., Hawanti, S., Aziez, F., Furqon, M., Zain, D. S. M., Anjarani, S., & Wu, T.-T. (2023). *The impact of AI chatbot-based learning on students' motivation in English writing classroom*. Porto, Portugal: The 6th International Conference on Innovative Technologies and Learning.

Sommet, N., Weissman, D. L., Cheutin, N., & Elliot, A. J. (2023). How many participants do I need to test an interaction? Conducting an appropriate power analysis and achieving sufficient power to detect an interaction. *Advances in Methods and Practices in Psychological Science, 6*(3). https://doi.org/10.1177/25152459231178728. Advance online publication.

* Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*, Article 1260843. https://doi.org/10.3389/fpsyg.2023.1260843.

* Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior, 160*, Article 108386. https://doi.org/10.1016/j.chb.2024.108386.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H. Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ, 366*, l4898. https://doi.org/10.1136/bmj.l4898

Stojanov, A. (2023). Learning with ChatGPT 3.5 as a more knowledgeable other: An autoethnographic study. *International Journal of Educational Technology in Higher Education, 20*(1), 35. https://doi.org/10.1186/s41239-023-00404-7

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, Article 100752. https://doi.org/10.1016/j.asw.2023.100752

Su, J., & Yang, W. (2023). Powerful or mediocre? Kindergarten teachers' perspectives on using ChatGPT in early childhood education. *Interactive learning environments*. Advance online publication. https://doi.org/10.1080/10494820.2023.2266490

* Suciati, S., Silitonga, L. M., Wiyaka, Huang, C.-Y., & Anggara, A. A. (2024). *Enhancing engagement and motivation in English writing through AI: The impact of ChatGPT-supported collaborative learning*. Tartu: International Conference on Innovative Technologies and Learning.

* Sun, D., Boudouaia, A., Zhu, C., & Li, Y. (2024). Would ChatGPT-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? An empirical study. *International Journal of Educational Technology in Higher Education, 21*(1), 14. https://doi.org/10.1186/s41239-024-00446-5.

* Svendsen, K., Askar, M., Umer, D., & Halvorsen, K. H. (2024). Short-term learning effect of ChatGPT on pharmacy students' learning. *Exploratory Research in Clinical and Social Pharmacy, 15*, Article 100478. https://doi.org/10.1016/j.rcsop.2024.100478.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.

Tai, T.-Y. (2024). Comparing the effects of intelligent personal assistant-human and human-human interactions on EFL learners' willingness to communicate beyond the classroom. *Computers & Education, 210*. https://doi.org/10.1016/j.compedu.2023.104965

Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences, 61*, 136–150. https://doi.org/10.1016/j.lindif.2017.11.015

Tan, C. N.-L., Tee, M., & Koay, K. Y. (2024). Discovering students' continuous intentions to use ChatGPT in higher education: A tale of two theories. *Asian education and development studies*. Advance online publication. https://doi.org/10.1108/AEDS-04-2024-0096

Titus, M. A. (2006). Detecting selection bias, using propensity score matching, and estimating treatment effects: An application to the private returns to a master's degree. *Research in Higher Education, 48*(4), 487–521. https://doi.org/10.1007/s11162-006-9034-3

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*(1), 15. https://doi.org/10.1186/s40561-023-00237-x

Tsai, C.-Y., Lin, Y.-T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12722-y
Tu, Y.-F., & Hwang, G.-J. (2023). University students' conceptions of ChatGPT-supported learning: A drawing and epistemic network analysis. *Interactive learning environments*. Advance online publication. https://doi.org/10.1080/10494820.2023.2286370
UNESCO Institute for Statistics. (2012). *International standard classification of education ISCED 2011*. UNESCO Institute for Statistics.
* Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education, 215*, Article 105031. https://doi.org/10.1016/j.compedu.2024.105031.
Urhan, S., Gençaslan, O., & Dost, Ş. (2024). An argumentation experience regarding concepts of calculus with ChatGPT. *Interactive learning environments*. Advance online publication. https://doi.org/10.1080/10494820.2024.2308093
Valcea, S., Hamdani, M. R., & Wang, S. (2024). Exploring the impact of ChatGPT on business school education: Prospects, boundaries, and paradoxes. *Journal of Management Education, 48*(5), 915–947. https://doi.org/10.1177/10525629241261313
van den Berg, G., & du Plessis, E. (2023). ChatGPT and generative AI: Possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. *Education Sciences, 13*(10), 998. https://doi.org/10.3390/educsci13100998
Vargas-Murillo, A. R., de la Asuncion Pari-Bedoya, I. N. M., & de Jesús Guevara-Soto, F. (2023). Challenges and opportunities of AI-assisted learning: A systematic literature review on the impact of ChatGPT usage in higher education. *International Journal of Learning, Teaching and Educational Research, 22*(7), 122–135. https://doi.org/10.26803/ijlter.22.7.7
Vázquez-Cano, E., Ramírez-Hurtado, J. M., Sáez-López, J. M., & López-Meneses, E. (2023). ChatGPT: The brightest student in the class. *Thinking Skills and Creativity, 49*, Article 101380. https://doi.org/10.1016/j.tsc.2023.101380
Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*, 3–14. https://doi.org/10.1007/s11409-006-6893-0
* Wahba, F., Ajlouni, A. O., & Abumosa, M. A. (2024). The impact of ChatGPT-based learning statistics on undergraduates' statistical reasoning and attitudes toward statistics. *Eurasia Journal of Mathematics, Science and Technology Education, 20*(7), 1–14. https://doi.org/10.29333/ejmste/14726.
Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education, 21*(1), 15. https://doi.org/10.1186/s41239-024-00448-3
Waltzer, T., Pilegard, C., & Heyman, G. D. (2024). Can you spot the bot? Identifying AI-generated writing in college essays. *International Journal for Educational Integrity, 20*(1), 11. https://doi.org/10.1007/s40979-024-00158-3
* Wang, X., & Feng, Y. (2023). *An experimental study of ChatGPT-assisted improvement of Chinese college students' English reading skills: A case study of dear life*. Barcelona, Spain: The 15th International Conference on Education Technology and Computers.
* Wang, X., Zhong, Y., Huang, C., & Huang, X. (2024). ChatPRCS: A personalized support system for English reading comprehension based on ChatGPT. *IEEE Transactions on Learning Technologies, 17*, 1762–1776. https://doi.org/10.1109/tlt.2024.3405747.
* Wiboolyasarin, W., Wiboolyasarin, K., Suwanwihok, K., Jinowat, N., & Muenjanchoey, R. (2024). Synergizing collaborative writing and AI feedback: An investigation into enhancing L2 writing proficiency in wiki-based environments. *Computers & Education: Artificial Intelligence, 6*, Article 100228. https://doi.org/10.1016/j.caeai.2024.100228.
Wijaya, T. T., Su, M., Cao, Y., Weinhandl, R., & Houghton, T. (2024). Examining Chinese preservice mathematics teachers' adoption of AI chatbots for learning: Unpacking perspectives through the UTAUT2 model. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12837-2
Williams, A. (2024). Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences. *International Journal of Educational Technology in Higher Education, 21*(1), 52. https://doi.org/10.1186/s41239-024-00485-y
Wise, B., Emerson, L., Van Luyn, A., Dyson, B., Bjork, C., & Thomas, S. E. (2024). A scholarly dialogue: Writing scholarship, authorship, academic integrity and the challenges of AI. *Higher Education Research and Development, 43*(3), 578–590. https://doi.org/10.1080/07294360.2023.2280195
Wong, R. M., & Adesope, O. O. (2021). Meta-analysis of emotional designs in multimedia learning: A replication and extension study. *Educational Psychology Review, 33*, 357–385. https://doi.org/10.1007/s10648-020-09545-x
Wong, L. H., Park, H., & Looi, C. K. (2024). From hype to insight: Exploring ChatGPT's early footprint in education via altmetrics and bibliometrics. *Journal of Computer Assisted Learning, 40*, 1428–1446. https://doi.org/10.1111/jcal.12962
Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12819-4
* Wu, C., Chen, L., Han, M., Li, Z., Yang, N., & Yu, C. (2024). Application of ChatGPT-based blended medical teaching in clinical education of hepatobiliary surgery. *Medical Teacher, Advance online publication*. https://doi.org/10.1080/0142159x.2024.2339412.
* Wu, T.-T., Lee, H.-Y., Li, P.-H., Huang, C.-N., & Huang, Y.-M. (2023). Promoting self-regulation progress and knowledge construction in blended learning via ChatGPT-based learning aid. *Journal of Educational Computing Research, 61*(8), 3–31. https://doi.org/10.1177/07356331231191125.
* Xiao, Q. (2024). ChatGPT as an artificial intelligence (AI) writing assistant for EFL learners: An exploratory study of its effects on English writing proficiency. *The 9th international conference on information and education innovations, verbania*.
* Xue, Y., Chen, H., Bai, G. R., Tairas, R., & Huang, Y. (2024). *Does chatgpt help with introductory programming? An experiment of students using ChatGPT in CS1 the 46th international conference on software engineering*. Lisbon: Software Engineering Education and Training.
Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies, 28*(11), 13943–13967. https://doi.org/10.1007/s10639-023-11742-4
Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370
Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. *System, 124*, Article 103351. https://doi.org/10.1016/j.system.2024.103351
Yeh, H.-C. (2024). The synergy of generative AI and inquiry-based learning: Transforming the landscape of English teaching and learning. *Interactive Learning Environments, Advance online publication*. https://doi.org/10.1080/10494820.2024.2335491
* Yilmaz, R., & Karaoglan Yilmaz, F. G. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers & Education: Artificial Intelligence, 4*, Article 100147. https://doi.org/10.1016/j.caeai.2023.100147.
Yun, W. S., & Surianshah, S. (2024). A review of the scholarly works on ChatGPT use in education: Bibliometric analysis. *International Journal of Technology in Education, 7*(3), 650–666. https://doi.org/10.46328/ijte.823
Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education, 21*(1), 21. https://doi.org/10.1186/s41239-024-00453-6
Zhai, X., Nyaaba, M., & Ma, W. (2024). *Can generative AI and ChatGPT outperform humans on cognitive-demanding problem-solving tasks in science? Science & education*. Advance online publication. https://doi.org/10.1007/s11191-024-00496-1
Zhai, C., & Wibowo, S. (2023). A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university. *Computers & Education: Artificial Intelligence, 4*, Article 100134. https://doi.org/10.1016/j.caeai.2023.100134
Zhan, Y., Yan, Z., Wan, Z. H., Wang, X., Zeng, Y., Yang, M., & Yang, L. (2023). Effects of online peer assessment on higher-order thinking: A meta-analysis. *British Journal of Educational Technology, 54*(4), 817–835. https://doi.org/10.1111/bjet.13310
* Zhang, J., Liu, Y., Cai, W., Wu, L., Peng, Y., Yu, J., Qi, S., Long, T., & Ge, B. (2024). *Investigation of the effectiveness of applying ChatGPT in dialogic teaching of electronic information using electroencephalography*. Xi'an: The 6th International Conference on Computer Science and Technologies in Education.
Zhang, P., & Tur, G. (2023). A systematic review of ChatGPT use in K-12 education. *European Journal of Education, 59*(2), Article e12599. https://doi.org/10.1111/ejed.12599

Zhao, X., Cox, A., & Cai, L. (2024). ChatGPT and the digitisation of writing. *Humanities and Social Sciences Communications, 11*, 482. https://doi.org/10.1057/s41599-024-02904-x

* Zhou, W., & Kim, Y. (2024). Innovative music education: An empirical assessment of ChatGPT-4's impact on student learning experiences. *Education and information technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-12705-z.

Zou, M., & Huang, L. (2023). The impact of ChatGPT on L2 writing and expected responses: Voice from doctoral students. *Education and Information Technologies, 29*, 13201–13219. https://doi.org/10.1007/s10639-023-12397-x