Journal Pre-proof

Color Me Confounded: A Critical Analysis of Media Comparisons on ChatGPT in Education

Alyssa P. Lawson, Amedee Marchand Martella, Joshua Weidlich, Miriam Mulders, Josef Buchner

PII: S0360-1315(26)00140-5

DOI: https://doi.org/10.1016/j.compedu.2026.105701

Reference: CAE 105701

To appear in: *Computers & Education*

Received Date: 19 June 2025

Revised Date: 9 June 2026

Accepted Date: 24 June 2026

Please cite this article as: Lawson A.P., Martella A.M., Weidlich J., Mulders M. & Buchner J., Color Me Confounded: A Critical Analysis of Media Comparisons on ChatGPT in Education, *Computers & Education*, https://doi.org/10.1016/j.compedu.2026.105701.

Color Me Confounded: A Critical Analysis of Media Comparisons on ChatGPT in Education

Alyssa P. Lawson[1,2], Amedee Marchand Martella[3], Joshua Weidlich[4,5], Miriam Mulders[6], and Josef Buchner[7]

1. Landmark College, United States
2. Colby College, United States
3. University of Georgia, United States
4. University of Zurich, Switzerland
5. Zurich University of Teacher Education, Switzerland
6. University of Duisburg-Essen, Germany
7. St. Gallen University of Teacher Education, Switzerland

Corresponding Author: Alyssa P. Lawson
Psychology Department
Colby College
805-453-2156
alyssaplawson@gmail.com

Declarations: No known conflicts of interest.

Use of Generative AI Tools: No Generative AI tools were used in analyzing or writing this manuscript. Generative AI was only used to help identify grammatical errors and improve clarity of several sentences.

Data Availability: Data will be made available upon request to the corresponding author.

CHATGPT CRITICAL ANALYSIS

Abstract

Research on the use of artificial intelligence in education is rapidly accumulating. In the ChatGPT literature, two recent meta-analyses were published (Deng et al., 2025 and Wang & Fan, 2025) that examined media comparisons between ChatGPT and more traditional forms of instruction; both reviews found ChatGPT to improve learning. However, as researchers have found in media comparisons studies (e.g., Clark, 1983; Lawson et al., 2024), instructional methods are often not controlled between conditions, making it difficult to attribute any learning differences to affordances of the medium itself. Therefore, the purpose of our critical review was to examine those media comparisons studies included in Deng et al. (2025) and Wang and Fan (2025) to understand whether the ChatGPT and control conditions were comparable on their instructional features (i.e., matched on instructional methods, practice with the dependent measure, and time spent learning the content). Results indicated a lack of control in the ChatGPT studies that were reviewed, with many studies conflating the use of ChatGPT with other instructional features. Further, across the various direct learning outcomes, studies often did not include enough information to determine the comparability of conditions. For the most common learning outcome (i.e., ChatGPT > Control), comparisons often involved a confound between conditions as well as missing information about instructional features. Therefore, the benefits of ChatGPT must be interpreted with caution at the present time, and more research is needed to determine how, for whom, and under what pedagogical conditions ChatGPT can improve learning.

Keywords: ChatGPT; Generative Artificial Intelligence; Media Comparisons; Methodological Rigor; Student Learning

Color Me Confounded: A Critical Analysis of Media Comparisons on ChatGPT in Education

Overview

The rapid proliferation of generative artificial intelligence (AI) tools, most prominently ChatGPT, has generated considerable excitement in research and educational practice alike. Enthusiasm is driven by the potential of ChatGPT to transform learning and teaching through personalized, adaptive interactions with learners (e.g., see Giannakos et al., 2024). For example, in a recent meta-analysis by Deng et al. (2025), which synthesized findings from 62 experimental studies, ChatGPT interventions were found to generally enhance academic performance, positively influence affective-motivational states, and facilitate higher-order thinking while reducing cognitive load. However, alongside the climate of excitement, researchers should remain wary of the robustness of empirical claims made about these tools. Given the historical context of educational technology research, marked by recurring cycles of hype and subsequent disillusionment (Mishra et al., 2009; Ramsey & West, 2023; Reich, 2020), the methodological and conceptual underpinnings of current research into the effectiveness of ChatGPT on learning remains central to investigate.

In their meta-analysis, Deng et al. (2025) themselves acknowledged significant methodological shortcomings in the included studies, such as inadequate sample size determination, lack of rigorous baseline controls, and simplistic measures of learning outcomes. In a critical commentary, Weidlich et al. (2025) questioned the soundness of the causal claims made by Deng et al. (2025), delineating three considerations that should be the minimum for any causal claims in this area: (1) are precise descriptions of the treatment provided? (2) is the control condition meaningful? and (3) are the outcome measures valid as indications of learning?

It is highly plausible that these same considerations are relevant to an even more recent meta-analysis on ChatGPT published by Wang and Fan (2025) [1], who use a similar approach and reported comparable findings. From 51 experimental studies, they concluded considerable benefits of ChatGPT on learning performance and moderate effects on learning perceptions and higher-order thinking.

While Weidlich et al. (2025) emphasized the importance of attending to three base considerations for causal claims, they did not present a critical analysis of whether the studies included in the Deng et al. (2025) and Wang and Fan (2025) meta-analyses accounted for factors that are relevant for a controlled (versus confounded) experimental comparison of ChatGPT conditions to control conditions. This question of control versus confounding illustrates a broader trend in the literature: meta-analyses are increasingly used to make claims about the educational effectiveness of novel technologies, yet often without adequately considering the comparability of experimental and control conditions across studies.

Expanding on Weidlich et al. (2025) and building on the methodology of estimating comparability versus confounding of immersive virtual reality (IVR) media comparisons by Lawson et al. (2024), this paper critically analyzed the primary studies included in the meta-analyses conducted by Deng et al. (2025) and Wang and Fan (2025). The goal of this analysis was to assess the comparability of experimental and control conditions from articles included in these influential meta-analyses of ChatGPT research with respect to key instructional features. The extent to which these conditions are comparable allows us to better understand if studies are well-controlled enough to draw specific conclusions regarding whether ChatGPT itself has a

[1] This meta-analysis was retracted on April 22, 2026 due to editor concerns about discrepancies that decreased confidence in the analyses, results, and conclusions. Despite this retraction, the individual studies on which it was based remain important to investigate as they reflect current empirical work on ChatGPT.

meaningful impact on learning or whether it is an opaque mix of ChatGPT and other instructional features that should be credited for the impacts on learning.

Media Comparison Research in Educational Technology

The use of media comparison research in educational technology has long been controversial, primarily due to the persistent methodological challenge of avoiding confounding the medium (or technology of interest) and the instructional methods embedded in the medium, commonly referred to as *media versus method*. Clark (1983) originally argued that instructional media do not inherently influence learning outcomes and that observed learning effects attributed to a medium frequently stem from underlying differences in instructional methods or content rather than from unique media properties. In response, Kozma (1991) argued that media possess distinct affordances that, when combined effectively with instructional methods, can indeed impact learning. Despite these differences, both authors underscored the need for using rigorous methodological approaches to consider the complexity of media- and technology-enriched learning and instruction (Clark, 1994; Hastings & Tracey, 2005; Kozma, 1994).

Nevertheless, this methodological rigor has often been absent in research practice. Recent systematic reviews indicate that media comparison studies regularly suffer from significant confounding. For instance, Buchner and Kerres (2023) found that the majority (80%) of studies examining augmented reality in education relied on direct media comparisons without sufficient methodological control, effectively obscuring instructional mechanisms underlying observed effects. Similarly, Lawson et al. (2024) reviewed IVR studies in STEM education, reporting that only 26% of the comparisons controlled instructional methods and content adequately, while 40% introduced significant confounding factors.

In another analysis of educational technology studies, Honebein and Reigeluth (2021) showed that media comparison studies are often masked as instructional comparisons (i.e., like a comparison of video instruction with traditional instruction). However, such comparative designs are always media-oriented as researchers compare a novel technology, like AI, to so-called traditional or conventional instruction based on commonly used media or technology (e.g., books).

The Promise of ChatGPT for Learning

Generative AI tools, such as ChatGPT, promise unprecedented personalization, scalability, and accessibility of learning experiences (Kasneci et al., 2023; Yan et al., 2024). Researchers and practitioners anticipate that these tools may fundamentally enhance feedback quality, individualize instruction, or support learners in ways previously only achievable through intensive human tutoring (Zhai, 2023). Despite such potential, ChatGPT is primarily a technological tool and, as such, it is without many unique inherent pedagogical properties. As Clark (1983) famously argued with respect to media, ChatGPT itself may not improve learning; rather, its educational effectiveness may depend on how it is employed for learning purposes (Bastani et al., 2024; Lehman et al., 2024).

Already, the instructional uses of ChatGPT vary dramatically. For instance, some studies employ ChatGPT merely as a proofreading or grammar-correction tool, while others integrate it as a sophisticated writing editor offering structural guidance and critical reflection prompts (e.g., Mahapatra, 2024; Niloy et al., 2023). ChatGPT has also been implemented as a personalized tutoring system, dynamically responding to learner inputs and actively guiding student engagement through iterative, conversational exchanges (e.g., Wu et al., 2024). Clearly, these diverse instructional integrations entail different expectations regarding learning outcomes. Such

use-specific effects of ChatGPT have been reported by Lehman et al. (2024), who found that students learned less when asking the system for solutions compared to when students asked for explanations. The former can lead to detrimental cognitive offloading (Stadler et al., 2024), where ChatGPT is being used as a crutch by students. In contrast, carefully designing AI-support scaffolding can yield durable learning gains (Yan et al., 2024). Critically, given the wide range of instructional uses of ChatGPT, the system itself does not need to change to yield widely diverging learning effects. In other words, careful scrutiny of the specific instructional implementation should underpin claims of ChatGPT's effectiveness for learning.

Contemporary Research on ChatGPT in Education

Reflecting on the excitement surrounding ChatGPT in education, Deng et al.'s (2025) and Wang and Fan's (2025) recently published meta-analyses provide early empirical evidence on ChatGPT in educational contexts. Based on 62 primary studies (although 69 were included in a systematic review of the literature), Deng and colleagues concluded that ChatGPT had a substantial positive impact on student learning and other desirable outcomes, reporting on overall effect size of g = .70. Wang and Fan identified 51 relevant primary studies, determining an effect on learning performance of g = .87. Taken at face value, these effect size estimates would make ChatGPT more effective than Intelligent Tutoring Systems (e.g. see Kulik & Fletcher, 2016), whose sole purpose is supporting learning.

However, these optimistic findings should be approached with caution, given the critical methodological issues highlighted in recent discussions. As previously mentioned, Weidlich et al. (2025) provided a critical commentary on Deng et al.'s meta-analysis, emphasizing three essential conceptual and methodological considerations required to interpret the effectiveness of generative AI interventions in education meaningfully: (1) clearly specifying the instructional

treatment; (2) defining a meaningful control condition; and (3) ensuring the validity of outcome measures as genuine indicators of student learning rather than task-specific performance differences or transient performance boosts from using AI systems.

With regard to this third consideration, we want to point out that both Deng et al. (2025) and Wang and Fan (2025) labeled their primary variables as 'academic performance' and 'learning performance,' respectively, instead of 'learning.' At first glance, this may guard against criticism of the validity of their studies' outcome variables; after all, performance measures could reflect students doing well simply because they are using ChatGPT, not because they have truly learned. However, we argue that this does not immunize them against this point of critique. In Deng et al.'s case, their title "Does ChatGPT enhance learning?" reveals a clear focus on learning itself, not just performance on tasks. Similarly, Wang and Fan recommended broad, grade-wide implementation of ChatGPT "to support student learning" (p. 17), a recommendation that would be on shaky ground if their results only showed task improvement. Finally, it is predictable that giving students a tool during assessments will boost their performance, much like how open-book exams typically yield higher scores. That kind of gain says little about actual learning gains made by students, and thus would not warrant a framing so closely tied to learning. Further, there is a critical difference between claiming that "ChatGPT is effective for learning" and recognizing that "ChatGPT can be implemented into learning environments to support learning." The former implies that the tool itself inherently causes learning, whereas the latter emphasizes the role of thoughtful instructional design alongside technology in shaping learning outcomes.

Importantly, although Weidlich et al. (2025) provided conceptual guidance on important considerations for conducting media comparison research, they did not quantify how well the

primary studies in both meta-analyses actually met these minimum methodological standards (as this is outside the scope of a commentary). Thus, while some primary studies were discussed as illustrations, a rigorous analysis is needed to ground the discussion more robustly in evidence. Moreover, additional analyses are required to evaluate whether studies in the meta-analyses were well-controlled or confounded. For example, even if both treatment and control conditions were adequately defined and specified, they could still diverge such that both instructional methods *and* ChatGPT were different between conditions. In this case the comparison would be confounded, meaning that any observed effects could not be confidently and solely ascribed to ChatGPT. Other potential confounding sources are the time spent within each treatment, or whether one group had more practice or experience with the outcome measure, among other sources (Lawson et al., 2024).

Present Research Goals

Given the rapid proliferation of ChatGPT-based interventions in education, conducting such a critical methodological evaluation can yield insights about the validity of claims regarding ChatGPT's specific effectiveness as well as empirically grounded guidance for designing future studies that avoid historical pitfalls associated with media comparison research (e.g., Buchner & Kerres, 2023; Lawson et al., 2024). To achieve these insights, our analysis followed two goals.

The first goal was to assess the primary studies in Deng et al. (2025) and Wang and Fan (2025) with respect to the comparability of experimental and control conditions *above and beyond* the nominal independent variable *ChatGPT*. That is, our analysis aimed to establish whether key instructional features were controlled (i.e., the same between conditions) or confounded (i.e., different between conditions). This goal traces back to Clark's (1983) key insights that we cannot determine the instructional effectiveness of a medium (or technology like

ChatGPT) when the comparison is confounded. Thus, our first research question (RQ1) was: Are the experimental and control conditions in the ChatGPT comparison studies analyzed by Deng et al. (2025) and Wang and Fan (2025) comparable on their instructional features (i.e., matched on instructional methods, practice with the dependent measure, and time spent learning the content)?

The second goal of this work was focused on the subset of comparisons that involved a direct measure of learning, as only these measures are suitable to demonstrate learning gains due to the ChatGPT treatment. For these comparisons, we aimed to explore, descriptively, what percentage of studies finding ChatGPT to be effective for learning (versus ineffective) tended to have conditions that were controlled or confounded on critical instructional features. Therefore, our second research question (RQ2) was: When looking at studies with direct learning outcomes, how often are ChatGPT and control conditions deemed comparable?

Although research on the impact of ChatGPT on learning is more expansive than that which is presented in these two meta-analyses, we decided to examine those studies in two prominent meta-analyses given that these meta-analyses have already shaped early claims about the effectiveness of generative AI in education. Analyzing the comparability of ChatGPT and control conditions is critical for drawing appropriate conclusions about the impact of ChatGPT interventions.

Method

Search Strategy

Deng et al. (2025)

Deng et al. (2025) included 69 articles published between the years 2022 (December) and 2024 (August) in their review of the effects of ChatGPT on student learning; 62 of these 69

articles were meta-analyzed. To be included in their review, studies needed to have (a) "used ChatGPT in the intervention;" (b) "used experimental and quasi-experimental designs;" (c) "used students as participants;" (d) "included at least one control group that did not use ChatGPT (or ChatGPT-supported learning applications) and one experimental group that did;" (e) "investigated the impact of ChatGPT on cognitive (e.g., knowledge acquisition), emotional (e.g., enjoyment), and psychological outcomes (e.g., self-efficacy);" (f) "explored the differential impact of ChatGPT across various age groups and educational levels;" (g) "restricted to peer-reviewed journal articles and conference papers;" (h) "a publication date of December 2022 or later;" and (i) been "published in English" (Deng et al., 2025, p. 7). The complete search strategy can be found starting on page 6 of their article.

Wang and Fan (2025)

Wang and Fan (2025) included 51 articles published between the years 2022 (November) and 2025 (February) in their review of the effects of ChatGPT on student learning. To be included in their review, as described in their Table 1, (a) "the research topics must be related to ChatGPT and learning performance, learning perception and higher-order thinking" and (b) "the research is written in English" (Wang & Fan, 2025, p. 4). To be excluded in their review, as described in their Table 1, the study (a) is "not an experimental or quasi-experimental study;" (b) does not have "experimental group with ChatGPT, control group without ChatGPT;" and (c) "does not contain data suitable for meta-analysis (e.g., Mean, SD, sample size, Cohen's *d, t* value)" (Wang & Fan, 2025, p. 4). The complete search strategy can be found starting on page 4 of their article.

Coding Details

Data Organization

The coding procedure of the present review took place using an Excel spreadsheet and followed a similar procedure used by Lawson et al. (2024). Each column represented a coding category, and each row represented a comparison between the control and experimental condition for each article. The experimental condition involved ChatGPT and the control condition did not involve the use of the AI tool.

Deng et al. (2025). Although there were 69 articles as part of Deng et al.'s main coding set, there were 70 rows (i.e., comparisons between a ChatGPT condition and a control condition) in our spreadsheet due to one article (Hu, 2024b) having two control conditions and one experimental condition. Therefore, control condition 1 versus experimental condition was represented in one row and control condition 2 versus experimental condition was represented in another row. As such, the spreadsheet had 70 rows reflecting 70 comparisons within 69 articles.

Wang and Fan (2025). Although there were 51 articles as part of Wang and Fan's main coding set, there was one article that had two control conditions (Roganović, 2024). Therefore, control condition 1 versus experimental condition was represented in one row and control condition 2 versus experimental condition was represented in another row. As such, the spreadsheet had 52 rows reflecting 52 comparisons within the 51 articles.

Development of Coding Categories

Our review was inspired by and generally patterned after the systematic review by Lawson et al. (2024) who examined the comparability of IVR conditions to more traditional forms of instruction (e.g., lectures, videos). Their coding scheme for determining the degree to which conditions were controlled on instructional methods and content was used as an initial

starting point for coding the articles in the present review. This coding scheme included five categories: (a) matched activities (were participatory activities controlled between conditions?); (b) matched non-IVR activities (were any activities that occurred outside of IVR controlled between conditions?); (c) matched practice with the dependent measure (were any practice opportunities that were related to the dependent measure controlled between conditions?); (d) matched time spent learning the content (was the lesson length controlled between conditions?); (e) and matched content (was the content of the lesson controlled between conditions?). Lawson et al. (2024) also included a category for operational definitions of the control and experimental conditions which was retained for our review (e.g., see also Martella et al., 2023, 2026). Finally, an additional coding category was added based on the fact that Deng et al. (2025) looked at different dimensions of student learning, including cognitive, emotional, and psychological outcomes, for a "holistic understanding of how this technology influences various aspects of student learning" (p. 7) and Wang and Fan (2025) looked at the impact of ChatGPT in promoting "student learning performance, learning perception, and higher-order thinking" (p. 2).

As part of this focus, across the two meta-analyses, the effects of ChatGPT were examined for academic performance, affective-motivational states, higher-order thinking, self-efficacy, and mental effort. However, the research team of the present review adopted a more stringent definition of learning for their second research question. This definition only included outcomes that were assessed on direct measures of learning (e.g., academic tests, exercises, projects/tasks with a scoring rubric) rather than indirect measures (e.g., perceptions of learning) or psychological/emotional measures such as motivation or self-efficacy. As such, we added a category to examine how many comparisons involved a direct dependent measure of learning.

CHATGPT CRITICAL ANALYSIS

The development of coding categories occurred using a subset of the 70 comparisons (and 69 articles) included in Deng et al.'s meta-analysis as this meta-analysis was published prior to Wang and Fan (2025). Training on the initial six categories (with slight modification by replacing "IVR" with "ChatGPT") occurred for all authors across five comparisons. After this training period, three authors independently double coded a combined total of 31 comparisons; the percent agreement was 60.33%. Because of the low agreement, the team met to refine the categories based on where discrepancies tended to occur. There were two main categories in which discrepancies tended to occur. The first problematic category was "matched non-ChatGPT activities" (modeled from "matched non-IVR activities" in Lawson et al., 2024). In the IVR literature as reviewed by Lawson et al. (2024), participants in certain IVR conditions would remove their headsets to participate in a real-world activity. As such, it was important to assess whether any activities that occurred outside of IVR were controlled between conditions such that the use of IVR, specifically, could be isolated. Given that ChatGPT is often used as a tool during a lesson activity and there is no "reality switching," this category was deemed not relevant for the present review.

The second problematic category was "matched content." Given that an inherent component of ChatGPT is producing content that can vary based on the prompts given and the questions asked and given that control conditions could involve the use of Google or other search engines to locate information, it was too difficult to determine the degree to which the content needed to match between conditions in order to be coded as "matched content." As such, this category was omitted from our review. Finally, the "matched activities" category was expanded to include other instructional methods (e.g., feedback, guided practice) that are not unique to ChatGPT and therefore should be controlled between conditions.

The initial training and coding procedure resulted in six final coding categories that were used to code the comparisons in both Deng et al. (2025) and Wang and Fan (2025). These categories included (a) operational definitions, (b) matched instructional methods, (c) matched practice with the dependent measure of learning, (d) matched time spent learning the content, (e) direct measure of learning, and (f) results of the study for direct measures of learning.

Given that determining what different instructional conditions entailed is predicated on sufficient operational definitions, the first category provided insight into the level of detail given for the control and experimental conditions. Categories 2-4 reflected instructional features that should be controlled between conditions to isolate the impact of using ChatGPT; these categories provided the needed information to answer the first research question about the comparability of ChatGPT and control conditions. Categories 5-6 provided the needed information to answer the second research question examining learning outcomes in relation to whether the conditions were comparable. Each of these categories and subsequent codes across both meta-analyses will be discussed below and are shown in Table 1, followed by a discussion of the coding procedure and interrater agreement.

Coding Categories and Subsequent Codes

Category 1: Operational Definitions. Providing an operational definition for instructional interventions is critical for understanding how conditions differ from one another and for determining which features have been isolated and/or controlled between conditions (Klahr, 2013; Lawson et al., 2024; Martella et al., 2023, 2026). To understand what occurred in the ChatGPT and control conditions, comparisons were assessed on two questions: (a) was an operational definition provided for the control condition? and (b) was an operational definition provided for the experimental condition? Codes included *yes*, *partial*, and *no*. To be coded as

yes, the description of the intervention needed to include all relevant information about what occurred in the condition such that it could be replicated by other researchers. To be coded as *partial*, the description of the intervention needed to involve more than a simple labeling of the condition (e.g., "lecture method," "traditional writing exercise") but not enough detail for replication to occur. To be coded as *no*, there needed to be a lack of description of the intervention, such as simply labeling the condition.

Category 2: Matched Instructional Methods. ChatGPT can be integrated into an educational lesson in a multitude of ways. For example, it can be used as a way to provide feedback to students, as a collaborative partner in solving problems, and as a questioning tool, among many other possibilities. The instructional conditions to which it can be compared can also vary substantially such as comparing the use of ChatGPT to the use of a Google search engine during a writing exercise or comparing the feedback provided via ChatGPT to more traditional instructor-given feedback during a math exercise. Key to being able to determine any *unique affordances* of using ChatGPT for learning purposes is ensuring all instructional methods not unique to ChatGPT are controlled between conditions. If the medium is confounded with the method, as is often the case with new educational technologies (e.g., see Clark, 1983), cause-and-effect conclusions about its impact become muddy (Lawson et al., 2024). Therefore, the coding team first listed the instructional features embedded in the conditions within the comparison and then assessed the comparison on the question: are the conditions comparable with regard to their instructional methods? Codes included *yes*, *no*, and *difficult to determine.*

To be coded as *yes*, any instructional methods not unique to ChatGPT needed to be controlled between conditions. For example, if ChatGPT was used as a feedback mechanism, the control condition also needed to receive some form of feedback. Or, if ChatGPT was used for

writing assistance, participants in the other condition needed to receive some form of writing assistance. Although there can be variation within each type of instructional method, we adopted a more liberal approach, choosing to focus on the alignment of the general method (e.g., did both conditions receive a form of feedback?) rather than all of the specific implementation procedures (e.g., was feedback implemented in exactly the same way between conditions?). It is important to note that any specific affordances of ChatGPT (e.g., personalized instruction/feedback, direct responses to inputted questions/prompts) were not considered confounds.

To be coded as *no*, any instructional methods not unique to ChatGPT needed to be confounded between conditions. For example, a confounded study would involve participants in the ChatGPT condition being afforded the opportunity to use ChatGPT to look up information during a problem-solving task while their peers in the control condition were tasked with solving the problems with no resources or assistance. The issue in this example is that there are two differences between conditions: the use of ChatGPT and problem-solving assistance. If the ChatGPT condition outperformed the control condition, the question becomes: "was it the use of ChatGPT for problem-solving assistance or simply the problem-solving assistance that resulted in greater learning gains?" Finally, if comparisons did not include enough information about the experimental and control conditions for the coding team to determine if the conditions were comparable, they were given a code of *difficult to determine*. Codes of *yes* reflected the conditions adhered to this instructional-feature control.

Category 3: Matched Practice with the Dependent Measure of Learning. Engaging in practice can be a powerful way to promote learning, particularly when the act of retrieving information from memory, or taking tests, can boost retention (Roediger & Karpicke, 2006, 2018). Exposing students to tasks or tests similar to those that will be used as a posttest during a

study can also familiarize them with the content and lead to a testing threat if not controlled between conditions (Lawson et al., 2024; Martella et al., 2013, 2026). To ensure it was the use of ChatGPT and not unequal practice with or exposure to the dependent measure that resulted in differential learning, both conditions should be controlled on this instructional feature. As such, comparisons were assessed on the question: did the conditions receive the same amount of practice with the dependent measure of learning? Codes included *yes*, *no*, *difficult to determine*, and *not applicable*.

To be coded as *yes*, participants in both conditions needed to receive or not receive learning opportunities that were similar to what was being assessed on the posttest. To be coded as *no*, participants in one condition needed to receive learning opportunities (e.g., tasks, tests) that were similar to what was being assessed on the posttest that participants in the other condition did not receive. For example, if participants are tested on their statistical reasoning and those in the ChatGPT condition receive explanations, justifications, and examples of choosing and using certain statistical tests during their practice tasks but their peers in the control condition do not receive this same type of practice, there would be unequal exposure to the dependent measure. If a comparison did not include enough information about what occurred during the intervention to determine if the conditions were controlled on this instructional feature, it was given a code of *difficult to determine*. Finally, given our focus on direct dependent measures of student learning, if comparisons did not directly measure learning, they would be coded as *not relevant* for matched practice with the dependent measure of learning. Codes of *yes* reflected the conditions adhered to this instructional-feature control.

Category 4: Matched Time Spent Learning the Content. Interventions can vary in their duration with some lasting for a short lesson (e.g., 30 min) and others spanning days,

weeks, or months. It is important to track the exposure to the intervention (also termed dosage) that participants receive (Mason & Smith, 2020) and ensure that the instructional time or time-on-task was consistent between conditions to rule out any confounding effects (Lawson et al., 2024; Martella et al., 2023, 2026). As noted by Clark (1983), one explanation for why participants may complete lessons using certain media in less time could be due to the increased effort and subsequent more effective lesson design given to the treatment condition. On the other hand, more instructional time may be positively related to student learning, although there are variable results in the literature (e.g., Anderson et al., 2016; Godwin et al., 2021; Wedel, 2021). Given these and other possibilities, ensuring participants in the ChatGPT condition receive the same amount of time spent learning the content as participants in the control condition can help to remove a dosage confound. Therefore, comparisons were assessed on the question "was time spent learning content controlled between conditions?" Initial codes included *yes*, *no*, *other*, and *difficult to determine*.

To be coded as *yes*, participants needed to receive the same amount of time to learn the content in both conditions (this code included if the average amount of time for both conditions was the same). To be coded as *no*, participants needed to receive different amounts of time to learn the content in both conditions. If a range of time or a maximum amount of time was listed, the comparison was coded as *other* for group discussion. After discussion, these codes were labeled either *likely yes* which indicated overlapping lesson lengths (e.g., 3-5 minutes versus 4-6 minutes) or the same maximum amount of time given to both conditions to complete the lesson (e.g., up to 1 hour) or *likely no* which indicated non-overlapping lesson lengths (e.g., 5-10 minutes versus 15-20 minutes) or different maximum amounts of time given to both conditions to complete the lesson (e.g., up to 1 hour versus up to 1.5 hours). Finally, if comparisons did not

include enough information to determine the specific amount of time participants spent learning the content in both conditions or the description was too general (e.g., both conditions had the intervention over 4 weeks), the comparison was given a code of *difficult to determine*. Codes of *yes* and *likely yes* reflected the conditions adhered to this instructional-feature control.

Category 5: Direct Measure of Learning. To determine whether comparisons involved a direct measure of learning, they were assessed on the question: "did the study involve a direct measure of learning?" Codes included *yes*, *no*, *difficult to determine*, and *discuss as a group*. To be coded as *yes*, the measure needed to go beyond self-report and capture student understanding/performance on tests, exercises, essays, or projects, for example. To be coded as *no*, the measure needed to be either self-report or a non-cognitive outcome such as motivation, engagement, or self-efficacy. If there was not enough information about the dependent measure to determine whether it was an independent assessment of learning, the comparison was coded as *difficult to determine*. Finally, if a coder wanted to discuss a grey area such as whether increases in creativity reflected "learning," they would code the comparison as *discuss as a group*. After discussion, the comparison would be given a primary code of *yes*, *no*, or *difficult to determine*. However, it is important to note that during some of these discussions over grey areas, the research team found that those in the ChatGPT condition were able to use the tool on the posttest. Given that the use of an aid during a posttest is a unique situation that should be highlighted, a comparison with this situation was given a special code of *No because ChatGPT was used during posttest*.

Category 6: Results of the Study for Direct Measures of Learning. To determine whether the ChatGPT condition resulted in greater performance on the direct measure of learning than the control condition in each comparison, the results of each study were coded as *ChatGPT*

> *Control*, *ChatGPT* < *Control*, *ChatGPT* = *Control*, *Mixed Findings*, *Inconclusive Statistics*, and *Not Applicable*. It is important to note that only direct measures of learning (discussed under Category 5) were examined. To be coded as *ChatGPT* > *Control*, the ChatGPT condition needed to result in statistically significantly greater performance on the dependent measure(s) of learning. To be coded as *ChatGPT* < *Control*, the control condition needed to result in statistically significantly greater performance on the dependent measure(s) of learning. To be coded as *ChatGPT* = *Control*, the conditions needed to result in non-statistically significant differences in performance on the dependent measure(s) of learning. *Mixed Findings* indicated that there was more than one direct measure of learning and that the results differed between/among these measures. If there were no inferential statistics provided or if there were any issues with the statistical analyses used to compare performance between the conditions (e.g., a paired samples *t*-test used to analyze a between-subjects research design), the comparison was coded as *Inconclusive Statistics*. Finally, if (a) a comparison did not have a direct measure of learning, (b)there was not enough information presented in the article to determine whether there was a direct measure of learning for the comparison (see Category 5), or (c) ChatGPT was used in the experimental condition during the posttest, the comparison was coded as *Not Applicable* for this category.

Coding Procedure and Interrater Agreement

The coding of comparisons included in Deng et al. (2025) occurred before the coding of comparisons included in Wang and Fan (2025) due to different publication dates. All articles were double coded to increase the rigor of the data extraction process. Percent agreement levels have been used to assess interrater agreement in previous work (Lawson et al., 2024; Martella et

al., 2026). Further, estimates using Krippendorff's Alpha (Hayes & Krippendorff, 2007) were also calculated for codes within each coding category.

Deng et al. (2025). The authors met to discuss coding of categories 1-5 and then proceeded to independently code five comparisons according to these categories. Questions and any discrepancies were discussed as a team before the authors were assigned their main coding set. Four of the authors were assigned 15 comparisons to independently code according to the first five coding categories as part of their main coding set. The first author independently coded all of these comparisons across these coding categories. The percent agreement was 83.10% across these categories which met the recommended interrater agreement level of at or above 80% (e.g., see Martella et al., 2013, 2026). Discrepancies were discussed and resolved by the research team. Finally, the first two authors independently coded all 70 comparisons according to the sixth and last coding category (results of the study for direct measures of learning); percent agreement was 82.86%. Discrepancies were discussed and resolved by these authors.

Wang and Fan (2025). Thirty-one of the 51 articles included in Wang and Fan's meta-analysis were also included in Deng et al.'s (2025) review. As such, these articles were already coded by the authorship team (as described above). For those 20 unique articles (and 21 comparisons), four articles were excluded from the coding procedure due to one article being a duplicate of another included article (Bašić et al., 2023a, 2023b), one article being retracted (Chan et al., 2024), one article not involving a comparison between an experimental and control condition (Lu et al., 2024), and one article being a meta-analysis (Wang et al., 2024). This left 16 articles (and 17 total comparisons) that were independently coded by the first two authors across the six coding categories. The percent agreement was 80.67% across these categories. Discrepancies were discussed and resolved by these authors.

Agreement Within Categories for Both Meta-Analyses. As previously mentioned, all articles across every coding category were coded by two independent coders. Interrater agreement for each category was calculated using percent agreement. Reliability was also assessed using Krippendorff's Alpha (Hayes & Krippendorff, 2007). Krippendorff's alpha was used to handle our categorical codes that included "difficult to determine" in addition to all other codes, such as "yes" or "no."

See Appendix A for Krippendorf's Alpha and interrater agreement percentages for each of the categories coded. Although some of the agreement levels and Krippendorf's Alpha values were below 80.00% agreement/alpha of .80, we want to highlight that every single article was double coded (i.e., was coded by two independent coders), and every discrepancy was discussed by the research team until agreement was reached. This process was done to ensure consistency of interpretations across the research team.

Data Analysis

Deng et al. (2025). Although all 69 articles (and 70 total comparisons) included in Deng et al.'s (2025) systematic review were coded[2], only the 62 articles they meta-analyzed were analyzed in our review given that these articles were used to make conclusions about ChatGPT's impact on student learning (broadly defined). One of these articles included the two comparisons (i.e., control condition 1 versus experimental condition and control condition 2 versus experimental condition). As such, codes were aggregated for each coding category across the 63 comparisons within the 62 articles.

[2] We did not receive information from Deng et al. on which seven articles were excluded from their meta-analysis until after our coding had concluded. Therefore, all 69 articles were coded, but we were able to analyze only those 62 articles included in their meta-analysis.

Wang and Fan (2025). Although 47 articles (and 48 total comparisons) included in Wang and Fan's (2025) meta-analysis were coded, two of the articles/comparisons were excluded from analyses. In these articles (Jafarian & Kramer, 2025; Karaman & Goksu, 2024), the participants did not interact with ChatGPT themselves (as was the case in all of the other articles included in their meta-analysis); rather, ChatGPT was used to design content that participants then interacted with as part of their lesson. Therefore, codes were aggregated for each coding category across 46 comparisons within 45 articles.

Results

Table 1 shows the results from each of the six coding categories for Deng et al. (2025) and Wang and Fan (2025). Results are divided according to the respective research question. Please note that results for each meta-analysis are presented separately and not aggregated due to many overlapping articles included in both meta-analyses.

Overview of Conditions

Category 1: Operational Definitions

Findings concerning operational definitions will be presented across conditions (with a representative example) as well as within each condition for the 63 comparisons in Deng et al. (2025) and the 46 comparisons in Wang and Fan (2025).

Across Both Conditions. In Deng et al (2025), the majority of comparisons (43 or 68.25%) involved either a partial or a full operational definition for both conditions. However, there were 20 comparisons (31.75%) that did not present at least a partial definition for both conditions. In Wang and Fan (2025), the majority of comparisons (32 or 69.57%) also involved either a partial or a full operational definition for both conditions. There were an additional 14

comparisons (30.43%) that did not present at least a partial definition for both the experimental and control groups.

Example. As an example of an article that included a comparison that involved at least a partial operational definition for each condition (and in this case, a full definition for both conditions), Darmawansah et al. (2025) provided a detailed figure of the experimental procedure for each condition (see Figure 4, p. 9), a description of the phases of ChatGPT use during the learning sessions (pp. 5-8) with sample scripts and interfaces, and explanations of the activities that occurred for both conditions during each week of the study (pp. 8-9).

ChatGPT Condition Only. In Deng et al (2025), approximately half of the comparisons (32 or 50.79%) included a ChatGPT condition that was fully operationalized and an additional 22 comparisons (34.92%) included a ChatGPT condition that was partially operationalized. A small portion of the comparisons (9 or 14.29%) had a ChatGPT condition that was not operationalized (either fully or partially). In Wang and Fan (2025), just over half of the comparisons (25 or 54.35%) included a fully operationalized description of the ChatGPT condition while an additional 14 comparisons (30.43%) included a partially operationalized description. A rather small portion of the articles did not provide any operationalized definition of the ChatGPT condition (7 or 15.22%).

Control Condition Only. In Deng et al (2025), there were fewer comparisons that provided a full operational definition of the control condition as compared to the ChatGPT condition. More specifically, over one-third of the comparisons included a fully operationalized definition (24 or 38.10%). An additional 19 comparisons (30.16%) provided a partial definition of the control condition. Finally, just under one-third of the comparisons (20 or 31.75%) had a control condition that was not operationalized (either fully or partially). In Wang and Fan (2025),

there were also fewer comparisons that provided a full operational definition of the control conditions (21 or 45.65%) as compared to the ChatGPT condition. Just under one-fourth of the articles provided a partial operationalized definition of the control condition (11 or 23.91%) and just under one-third of the comparisons (14 or 30.43%) did not include any operationalized definition of this condition.

RQ 1: Are the Experimental and Control Conditions in the ChatGPT Comparison Studies Analyzed by Deng et al. (2025) and Wang and Fan (2025) Comparable on Their Instructional Features?

Findings for each instructional-feature control will be presented first followed by the overall comparability of the conditions for the 63 comparisons in Deng et al. (2025) and the 46 comparisons in Wang and Fan (2025). Representative examples of each category are provided as well.

Category 2: Matched Instructional Methods

In Deng et al. (2025), there were only 11 comparisons (17.46%) that were deemed to have matched instructional methods between the ChatGPT and control conditions. Almost one-third of the comparisons (19 or 30.16%) were deemed to have ChatGPT and control conditions that were not matched on instructional methods. For these comparisons, 18 (94.74%) included instructional methods that were favorable toward the ChatGPT condition as compared to the control condition (e.g., the ChatGPT condition may have included feedback or problem-solving support that was not also provided to the control condition). Only one comparison (5.26%) included instructional methods that were favorable toward the control condition. Finally, over half of the comparisons (33 or 52.38%) did not involve enough information about the ChatGPT

and control conditions to make a determination about whether the instructional methods were matched.

In Wang and Fan (2025), even fewer of the articles were deemed to have matched instructional methods (4 or 8.70%). Over one-third of the comparisons (17 or 36.96%) were deemed to not have matched instructional methods between the two conditions. Of these 17 comparisons, 16 (94.12%) included instructional methods that were favorable toward the ChatGPT condition as compared to the control condition. One comparison (5.88%) was not clear on which group would benefit more from the intervention. Finally, over half of the comparisons (25 or 54.35%) did not involve enough information about one or both of the conditions to determine whether the instructional methods were matched.

Example. As an example of matched instructional methods, Zhang et al. (2024) had students in both conditions engage in dialogue-based instruction; this dialogue involved ChatGPT for the experimental condition and a real-life teacher for the control condition (see Figure 1 in their paper, p. 151). The teacher was trained on conducting dialogue-based teaching and ChatGPT was informed about the lecturer role it would take to facilitate learning. As such, participants in both conditions were able to ask questions to their teacher (either ChatGPT or a real-life person) and engage in a dialogue about histogram equalization.

Category 3: Matched Practice with the Dependent Measure

In Deng et al. (2025), there were 15 comparisons (23.81%) that did not involve a direct measure of learning and were coded *not relevant*. For those 48 comparisons that did directly measure learning (76.19%), 14 comparisons (29.17%) had conditions that were matched on any practice participants received with the dependent measure of learning. Only six comparisons (12.50%) included unequal practice with the dependent measure, exclusively benefiting the

ChatGPT condition. Finally, 28 comparisons (58.33%) did not involve enough information about the ChatGPT and control conditions to determine whether one condition had more practice with the dependent measure than the other condition.

In Wang and Fan (2025), there were 10 comparisons (21.74%) that did not involve a direct measure of learning and were coded *not relevant*. For the remaining 36 comparisons that did directly measure learning (78.26%), 15 comparisons (41.67%) had conditions that were matched on any practice participants received with the dependent measure of learning. Only four comparisons (11.11%) included unequal practice with the dependent measure. Finally, 17 comparisons (47.22%) did not involve enough information about the conditions to determine if there was an equal amount of practice with the dependent measure.

Example. As an example of matched practice with the dependent measure, Xiao (2024) compared a 10-week writing program that involved the use of ChatGPT to one that involved traditional writing instruction. They noted the same instructional materials and curriculum were used for both conditions (p.53) and outlined the experimental arrangement in their Table 1 (p. 54) which showed consistency in the learning focus and writing tasks provided to participants in both conditions. The dependent measure was a standardized writing test in which participants wrote an argumentative essay; the writing program involved in both conditions had the objective of helping students to develop the skills to write argumentative essays.

Category 4: Matched Time Spent Learning the Content

In Deng et al. (2025), just under one-third of the comparisons (19 or 30.16%) were matched on the time allotted for participants to learn the content. Additionally, two comparisons (3.17%) were coded as *likely yes* for time being matched. Only two comparisons (3.17%) were explicitly not matched on time, and an additional two comparisons (3.17%) were coded as *likely*

not matched. Finally, under two-thirds of the comparisons (38 or 60.32%) did not involve enough information to determine whether the conditions were matched on time spent learning the content.

In Wang and Fan (2025), just over one-fourth of the comparisons (13 or 28.26%) were matched on the time spent learning content across conditions. Additionally, three comparisons (6.52%) were coded as *likely yes* for time being matched. Only one comparison (2.17%) was coded as not matched on time and one more comparison (2.17%) was coded as *likely not* matched. Finally, under two-thirds of the comparisons (28 or 60.87%) did not involve enough information to determine whether the conditions were matched on the time spent on learning the content.

Example. As an example of matched time spent learning the content, Suciati et al. (2024) provided both conditions with 600 minutes (100 minutes per each of six meetings) in the experimental phase.

Overall Comparability of the Instructional Features

To determine the overall comparability of conditions based on their instructional features, the number of control issues were examined across Categories 2-4 (i.e., instructional methods used, practice with the dependent measure, and time spent learning the content). Two analyses were conducted (as done by Lawson et al., 2024): one in which benefit of the doubt was not given—meaning that any *difficult to determine* codes were counted as violating the control—and one in which benefit of the doubt was given—meaning that any *difficult to determine* codes were counted as *not* violating the control.

No Benefit of the Doubt. Figure 1 shows the percentage of the 63 comparisons from Deng et al. (2025) that violated these instructional-feature controls when comparisons were not

given the benefit of the doubt. Only six comparisons (9.52%; Darmawansah et al., 2025; Hu, 2024a; Hu, 2024b comparison 1; Song & Song, 2023; Stadler et al., 2024; and Suciati et al., 2024) adhered to all three instructional-feature controls. This finding means that almost all of the comparisons (57 or 90.48%) violated or had missing information for at least one of the three controls. The average number of violations/missing information was 2.00 (*SD* = .98) per comparison.

Figure 1 also shows the percentage of the 46 comparisons from Wang and Fan (2025) that violated these instructional-feature controls when comparisons were not given the benefit of the doubt. Only three comparisons (6.52%; Darmawansah et al., 2025; Hu 2024a; and Song & Song, 2023) adhered to the three instructional-feature controls. Therefore, the majority of articles (43 or 93.48%) violated or had missing information for at least one of the three controls. The average number of violations/missing information was 2.02 (*SD* = .83) per comparison.

Benefit of the Doubt. Figure 2 shows the percentage of the 63 comparisons from Deng et al. (2025) that violated the three instructional-feature controls when comparisons were given the benefit of the doubt. There were 43 comparisons (68.25%) that adhered to instructional-feature controls across Categories 2-4, and 20 comparisons (31.75%) that explicitly did not adhere to at least one of the three controls. The average number of violations was .46 (*SD* = .78) per comparison.

Figure 2 also shows the percentage of the 46 comparisons from Wang and Fan (2025) that violated the instructional-feature controls when comparisons were given the benefit of the doubt. There were 29 comparisons (63.04%) that adhered to instructional-feature controls across the three categories, and 17 comparisons (36.96%) that explicitly did not adhere to at least one of the three controls. The average number of violations was .50 (*SD* = .75) per comparison.

Figure 1 Versus Figure 2. The stark difference in results between the comparisons receiving or not receiving the benefit of the doubt across the two meta-analyses indicates how prevalent missing information was across the articles.

RQ2: When Looking at Studies with Direct Learning Outcomes, How Often Are ChatGPT and Control Conditions Deemed Comparable?

Category 5: Direct Measure of Learning

In Deng et al. (2025), just under one-fourth of the 63 comparisons (14 or 22.22%) involved learning that was not directly measured and that often took the form of self-report scales to assess participants' perceptions of learning, self-efficacy, and motivation (as reflected in the other outcomes of interest to Deng et al., 2025 that were generally referred to as "learning"). In one comparison (1.59%), there was not enough information to determine if and how learning was assessed. We identified learning as directly measured in 48 comparisons (76.19%); however, seven of the comparisons involved the use of ChatGPT during the posttest for the experimental condition. For example, in the study by Moneus and Al-Wasy (2024) examining the impact of ChatGPT on translation for students, participants in the experimental group used ChatGPT to translate the text during the test while the control group was instructed to translate the text manually, without any technological tools. The use of ChatGPT on the dependent measure is problematic given that the test does not truly assess unaided learning gains and does not afford a fair comparison between learning outcomes in both conditions. Therefore, only 41 comparisons (65.08%) were deemed to have *true* direct measures of learning.

In Wang and Fan (2025), just under one-fourth of the 46 comparisons (10 or 21.74%) involved learning that was not directly measured. Although we identified learning as directly measured in 36 comparisons (78.26%), eight of these comparisons involved the use of ChatGPT

during the posttest for the experimental condition. Therefore, only 28 comparisons (60.87%) were deemed to have *true* direct measures of learning.

Example. As an example of a direct measure of learning, Wu et al. (2024) had a posttest on the mathematics vector unit in which students participated that included 20 multiple-choice questions written by two mathematics experts.

Category 6: Results of the Study for Direct Measures of Learning

For those 41 comparisons examined in Deng et al. (2025) that involved a direct measure of learning, it was reported that the ChatGPT condition had significantly better learning outcomes than the control condition in 27 comparisons (65.85%). In two of the comparisons (4.88%), the control condition outperformed the ChatGPT condition. There was no difference between the ChatGPT and control conditions in nine of the comparisons (21.95%). In two of the comparisons (4.88%), the statistics were inconclusive, either because no inferential statistics were presented or because inappropriate statistics were used. Finally, one comparison (2.44%) had mixed findings across multiple learning outcomes.

For those 28 comparisons examined in Wang and Fan (2025) that did involve a true direct measure of learning, it was reported that the ChatGPT condition had significantly better learning outcomes than the control condition in 18 comparisons (64.29%). There was only one comparison (3.57%) in which the control condition outperformed the ChatGPT condition. There were an additional six comparisons (21.43%) where there was no significant difference between the two conditions. In two of the comparisons (7.14%), the results were mixed, with some measures of learning showing significant differences while others did not. Finally, there was one comparison (3.57%) that was inconclusive due to the lack of inferential statistics presented in the article.

Exploring the Comparability of ChatGPT and Control Conditions and the Results of Each Study

In service of answering our second research question, only those 41 comparisons in Deng et al. (2025) and 28 comparisons in Wang and Fan (2025) that had true direct measures of learning were examined. For these articles, we separated comparisons by the reported impact of ChatGPT relative to control conditions. For each outcome type identified, we examined the number of comparisons that were deemed comparable, not deemed comparable, or were labeled difficult to determine due to missing information. To be considered "comparable" for this specific analysis, the comparisons between the ChatGPT and control conditions had to explicitly adhere to all three instructional-feature controls discussed for RQ 1 (i.e., Categories 2-4). If it was not clear whether the comparison adhered to each of the controls due to missing information, it was labeled *difficult to determine*. See Figures 3 and 4 for the frequency of comparisons in which the ChatGPT and control conditions were deemed comparable, not comparable, or difficult to determine for the studies in the Deng et al. (2025) and Wang and Fan (2025) meta-analyses. It is important to acknowledge that the high proportion of "difficult to determine" cases identified in the results limits the strength of any pattern-based inferences regarding the relationship between confounding and learning gains.

ChatGPT Condition Better (ChatGPT > Control). For those 27 comparisons examined in Deng et al. (2025) in which the learning outcome demonstrated that the ChatGPT condition performed significantly better on the posttest than the control condition, only one of these comparisons (3.70%) had conditions that were considered comparable on their instructional features. Eleven of the 27 comparisons (40.74%) were not comparable. Finally, approximately half of the 27 comparisons (15 or 55.56%) were difficult to determine regarding their conditions'

comparability due to a lack of information provided for at least one of the instructional-feature controls.

For those 18 comparisons examined in Wang and Fan (2025) in which the learning outcome demonstrated that the ChatGPT condition performed significantly better on the posttest in comparison to the control condition, two comparisons (11.11%) were considered comparable on their instructional features. Six of the comparisons (33.33%) were considered not comparable. Lastly, more than half of the comparisons (10 or 55.56%) were difficult to determine regarding their conditions' comparability due to a lack of information provided for at least one of the instructional-feature controls

Control Condition Better (ChatGPT < Control). For those two comparisons examined in Deng et al. (2025) in which the learning outcome demonstrated that the control condition outperformed the ChatGPT condition on the posttest, one comparison (50.00%) had conditions that were comparable on their instructional features and one condition (50.00%) had conditions that were not comparable on these features.

For the one comparison examined in Wang and Fan (2025) in which the learning outcome demonstrated that the control condition had better performance on the posttest in comparison to the ChatGPT condition, it was difficult to determine whether the conditions were comparable due to a lack of information provided for at least one of the instructional-feature controls.

Tied (ChatGPT = Control). For those nine comparisons examined in Deng et al. (2025) in which the ChatGPT condition and the conventional condition were tied (i.e., did not have statistically significantly different results), zero comparisons (0.00%) had conditions that were comparable on its conditions' instructional features. Further, there were two comparisons

(22.22%) that had conditions that were not comparable. The majority of the nine comparisons (7 or 77.78%) were difficult to determine regarding their conditions' comparability due to a lack of information provided for at least one of the instructional-feature controls.

For those six comparisons examined in Wang and Fan (2025) in which there were no significant differences between the ChatGPT and the conventional conditions, zero comparisons (0.00%) had conditions that were comparable, and one comparison (16.67%) had conditions that were not comparable. The majority of the comparisons (5 or 83.33%) were difficult to determine whether the conditions were comparable.

Mixed Findings. For the one comparison examined in Deng et al. (2025) that had mixed findings across different learning outcomes, it was difficult to determine if the conditions were comparable due to a lack of information provided for at least one of the instructional-feature controls.

For the two comparisons examined in Wang and Fan (2025) in which there were mixed findings regarding the impact of ChatGPT on learning outcomes between conditions, one comparison (50.00%) had conditions that were not comparable while the other comparison (50.00%) had conditions that were difficult to determine with regard to their comparability.

Inconclusive Results. For those two comparisons examined in Deng et al. (2025) that were classified as inconclusive, both comparisons (100.00 %) were difficult to determine regarding their conditions' comparability due to a lack of information provided for at least one of the instructional-feature controls.

For the one comparison in Wang and Fan (2025) in which the findings were inconclusive, the comparison was difficult to determine regarding whether the conditions were comparable.

Discussion

Our critical review was guided by two primary research questions aimed at critically analyzing the ChatGPT studies included in both Deng et al.'s (2025) and Wang and Fan's (2025) meta-analyses that compared ChatGPT-supported learning to traditional instructional conditions. These research questions were investigated by (a) examining the extent to which ChatGPT and control conditions were comparable on their instructional features and (b) examining the percentage of comparisons that did or did not involve comparable conditions across various direct learning outcomes. We begin by synthesizing the main empirical patterns that emerged from our coding across the 63 ChatGPT media comparisons in Deng et al.'s meta-analysis and the 46 ChatGPT media comparisons in Wang and Fan's meta-analysis. In doing so, we critically examine the methodological quality of the included studies as they relate to key instructional features, connecting our findings to established debates in educational technology research. We also offer guidance for future inquiry.

Summary of Main Findings

Methodological Confounds are Prevalent

Our analysis revealed that a portion of the research studies on ChatGPT do not meet the investigated controls discussed in this research. Echoing long standing critiques in the media comparison literature (Buchner & Kerres, 2023; Clark, 1983; Lawson & Martella, 2023; Lawson et al., 2024; Martella et al., 2026), we found that a substantial share of studies conflated the use of ChatGPT with other instructional features that can be replicated in non-ChatGPT environments and that should be held constant between conditions. If this conflating is true beyond just the sample of studies we examined, it would be extremely difficult to understand the impact of ChatGPT, specifically, on learning. In the studies analyzed in the present review, only

a small fraction of comparisons between a ChatGPT condition and a control condition could be deemed controlled on three key instructional features: (1) matched instructional methods, (2) matched practice with the dependent measure of learning, and (3) matched learning time. This issue replicates and extends the concerns raised by Weidlich et al. (2025), who noted that many ChatGPT studies in Deng et al. (2025) failed to meet even minimal standards for experimental rigor.

As in prior media waves, the reviewed subset of studies demonstrated that ChatGPT is often introduced alongside enriched instructional features, such as providing feedback or generative prompts (Stadler et al., 2024). This confounding is problematic: ChatGPT is not *pedagogy* but a medium that *affords* certain pedagogical uses. Unless those pedagogies that are not unique to ChatGPT itself are mirrored or controlled in the comparison group, effects attributed to ChatGPT may actually reflect the instructional methods it was used to deliver. Therefore, trying to understand the unique contributions ChatGPT may provide in learning is difficult to determine, particularly if the pattern of results holds across a larger sample of ChatGPT research. Further, given the finding that there were some studies in which the ChatGPT condition received more practice with the dependent measure of learning, the treatment effect may be artificially inflated. Consistent with earlier findings in immersive media research (Lawson et al., 2024), practice with the dependent measure remains underreported and uncontrolled in much of the reviewed literature.

A particularly prevalent issue across the studies analyzed was the asymmetrical provision of scaffolding or feedback. More specifically, across a portion of these studies, ChatGPT served as a support/feedback mechanism while control groups received minimal or no equivalent support. Although scaffolding and feedback are strong capabilities that ChatGPT can provide,

the provision of this support can also occur in conventional conditions. If the goal of this research is to understand if ChatGPT itself enhances learning, these supports should be controlled between conditions in media comparison research. These findings are consistent with problems occurring across various instruction media (e.g., Clark, 1983), suggesting that benefits attributed to *technology* often derive from richer instructional features, not the medium itself. Similarly, in line with immersive media research (Lawson et al., 2024; Martella et al., 2026), we also found that many of these studies failed to report or control time-on-task, leaving open the possibility that ChatGPT groups simply engaged longer with the content or may have benefitted from a more efficient intervention. Our coding scheme was intentionally conservative with regard to the dosage variable, in service of reducing threats to internal validity. This control was a tradeoff in that it could have led to an underestimation of possible efficiency-related advantages of ChatGPT when learning time was not measured or controlled. To determine if time-on-task is a focal causal mechanism benefitting ChatGPT interventions, this dosage variable should be isolated and systematically studied in future studies.

Finally, in pursuit of our second research question, we examined the direct learning outcomes of each comparison and explored how often the ChatGPT and control conditions were deemed comparable across these learning outcomes. For direct learning outcomes, more than half of the studies reviewed from the two meta-analyses found ChatGPT conditions led to significantly better learning outcomes than control conditions (65.85% in Deng et al., 2025 and 64.29% in Wang & Fan, 2025).

When examining the comparability of conditions for various learning outcomes, it was difficult to gather a clear picture of what might be driving effects due to smaller sample sizes for some of the outcomes (e.g., Control > ChatGPT, mixed findings, inconclusive results) and due to

missing information that prevented an assessment of condition comparability. These issues substantially limit the strength of any inferences regarding the relationship between confounding and learning outcomes. Perhaps most problematic was the degree of missing information in these studies. Systematic and detailed research is needed to determine whether ChatGPT itself causes changes in learning, whether it is the instructional methods themselves, or whether it is the synergy among the methods and the AI medium.

As it stands, we can determine that a large percentage of the reviewed research finding ChatGPT to be more effective than control interventions (i.e., ChatGPT > Control) was coded as 'confounded' within our coding scheme. More specifically, 11 comparisons (40.74%) were not deemed to have comparable conditions in Deng et al. and six comparisons (33.33%) were not deemed to have comparable conditions in Wang and Fan. However, due to the missing critical methodological information, the comparability of the remaining 15 comparisons (55.56%) in Deng et al. and 10 comparisons (55.56%) in Wang and Fan was inconclusive. This issue of missing information was problematic across all direct learning outcomes, limiting the conclusions that can be drawn.

The Illusion of Precision in ChatGPT Studies: A Critical Appraisal

Our synthesis underscores a potential problem that needs to be further investigated: if the results of the reviewed literature are representative of the field as a whole, a large proportion of research on ChatGPT introduces it as a technological intervention without isolating the pedagogical mechanisms through which it operates. If this lack of isolation is generally the case, the comparison is less about *whether* ChatGPT works and more about *what else* in tangent with ChatGPT improved the learning environment. The current evidence base, provided in two meta-

analyses investigating the impact of ChatGPT on learning appears to be conceptually and methodological fragile.

One key problem identified in our analysis (that may be important in a broader context) concerns the operational definitions of both experimental and control conditions. It was not infrequent to find the term *ChatGPT-based instruction* used as a catch-all category, without specifying the nature of the instructional interaction (e.g., explanation versus completion support), the input format (e.g., free prompts versus structured tasks), or the role of the human instructor. Accordingly, a proportion of studies included in Deng et al. (2025) and Wang and Fan (2025) did not offer full operational definitions such that all critical aspects of the conditions were outlined, particularly for the control group. Without detailed procedural information, it is impossible to determine what participants actually experienced in each condition or to pinpoint how conditions were similar and how they were different from one another (e.g., see Klahr, 2013). The lack of transparency in the analyzed studies undermines reproducibility and comparability and calls into question the robustness of the reported effects. This issue is particularly concerning if the results of the present review are representative of the other relevant literature in the field.

Furthermore, another common issue we noted was when studies provide minimal information about the control condition, such as using the vague labels like *traditional instruction* or *the non-ChatGPT group*. Many ChatGPT conditions also suffered from a lack of specificity leaving the readers to assume what is meant by a group of students receiving a ChatGPT intervention. As found in other literature bases (e.g., active learning [Freeman et al., 2014; Martella et al., 2023] or immersive virtual reality [Lawson et al., 2024]), educational interventions can vary widely in how they are implemented. This variation in implementation

makes it difficult to aggregate effect sizes meaningfully in meta-analyses and to provide prescriptive guidance to instructors (Martella & Schneider, 2024). As others have noted (e.g., Lawson et al., 2024; Martella et al., 2023), precise operationalization of both conditions is indispensable for meaningful comparisons. Without knowing what control group participants actually did, claims about ChatGPT's superiority remain speculative.

Finally, practice or exposure to outcome measures also requires some scrutiny. Several studies in this review relied on post-intervention assessments that closely mirrored (in content and/or form) the learning tasks practiced in the ChatGPT condition but not in the control group. This practice introduces a possible testing threat (Roediger & Karpicke, 2006) and raises doubts about the generalizability of reported learning gains. If ChatGPT conditions allow for more targeted rehearsal of and/or exposure to elements of the outcome task, effects may reflect *task familiarity* and not deep learning. Additionally, given that there were seven comparisons in which the ChatGPT condition was afforded the opportunity to use the tool on the final assessment of learning in Deng et al. (2025) and eight comparisons in Wang and Fan (2025), we must question whether these studies are truly measuring student learning/understanding. Receiving aid on a final assessment in one condition and not in the other is a confound that undermines the conclusions that can be made about the effects of the independent variable.

More Than a Tool, Less Than a Treatment: Implications for Interpreting ChatGPT Effects

The reviewed evidence base does not seem to support unequivocal claims about the *causal* effectiveness of ChatGPT in education. While the reviewed studies often report statistically significant performance gains in ChatGPT conditions, the pervasive lack of methodological control and lack of detailed instructional information makes it difficult to know whether these changes should be attributed to the technology itself or to the instructional

practices being used alongside ChatGPT. This finding aligns with Clark's (1983) seminal argument that media do not cause learning and rather, it is the instructional methods that drive educational outcomes. Indeed, within the reviewed studies, we found that ChatGPT often embodied good instructional design (e.g., feedback, additional content/information) that can and should be replicated in more traditional instructional environments. But, to truly understand the impact ChatGPT itself can have on learning, it is vital to isolate the features that are unique to ChatGPT. Without such isolation, the tool's reported effectiveness may say more about how well it is integrated with pedagogy than about any inherent affordances of generative AI.

This disentanglement is important to deeply consider at all stages of research and dissemination, both within research on ChatGPT and more broadly in research on educational technology. If a researcher would like to know if a technology itself enhances learning and makes the claim that technology X is effective for learning, then the study should be designed to isolate the effects/affordances of technology X. This design would involve holding all instructional methods that are not unique to the technology constant between conditions such that causal conclusions can be attributed to the technology rather than to the methods (Lawson et al., 2024). However, if the study were designed such that both the technology and established instructional methods (e.g., feedback, generative learning strategies) that were embedded within it/included alongside it differed between conditions, researchers should not and cannot draw the conclusions that only the technology was effective for learning. The conclusion would be more accurate if stated as, "the instructional package involving technology X with the instructional methods A, B, and C led to greater learning."

Many studies have compared interventions that differ in many ways from one another with the goal of making general conclusions about whether, for example, human tutoring is more

effective than studying texts (e.g., see Vanlehn et al., 2007). These types of studies offer insight into whether a more business-as-usual condition might be better replaced by a new instructional procedure. However, what these studies do not allow for is the specific isolation of *why* certain interventions were more or less effective.

A goal of the educational technology literature is often to make specific claims about the unique contributions of specific technological tools. However, as found in this review, the studies do not always involve the control of instructional features (e.g., feedback, activities) between conditions to isolate the contributions of the technology itself. If the goal is to understand the *general* impact of an instructional package (i.e., intervention involving many instructional features), more than one difference in the instructional features between conditions may be appropriate as long as conclusions remain general and do not pinpoint one feature as the causal ingredient. If the goal is to target key instructional features of a new intervention as causal factors and to provide prescriptive guidance, variables should be isolated and controlled accordingly. Therefore, we recommend researchers ensure that the research design of their study aligns with their specific research question(s) and that they align their conclusions carefully with their methodological designs.

The results of our review also highlight variations in how ChatGPT is used across studies, ranging from proofreading assistance to a fully interactive tutor. This variation raises important questions about the conceptual coherence of research in this field. When the term *ChatGPT* can mean radically different things across interventions, generalizing the effects on learning becomes problematic. This inconsistency reflects a broader issue in emerging technology research: the tendency to evaluate tools without specifying their pedagogical functions (Reeves & Oh, 2017). In light of these findings, if the trends found here are consistent

across other relevant studies, it would be problematic to *currently* say that ChatGPT (as a treatment itself) directly affects learning. Future studies and reviews should focus on disentangling *what* instructional functions ChatGPT serves (e.g., explanation, motivation, regulation) and *how* these functions interact with learner characteristics and task demands. Only then can we meaningfully assess whether generative AI constitutes an educational advance or merely a novel form of delivery.

Toward Methodological Clarity and Pedagogical Precision: Recommendations for Future Research

Although the results found here are based on studies within two meta-analyses on ChatGPT, there are important takeaways that should be discussed. First, researchers should ensure they provide detailed operational definitions of both the experimental and the control conditions. These definitions include not only stating that ChatGPT was utilized but also specifying how it was embedded pedagogically—for example, what prompts were employed, what roles it fulfilled within the instructional sequence, and what expectations were placed on learners. This level of detail should also be included in describing the comparison condition. Without such detail, replication and interpretation remain limited.

Second, future studies should ensure equivalence in those instructional features not unique to ChatGPT across conditions if the unique contributions of ChatGPT are to be isolated. Instructional support mechanisms such as feedback, scaffolding, and guided practice should not be exclusive to the ChatGPT condition unless these differences are central to the research question and are identified as an instructional method not as a specific and unique feature of the technology. If left uncontrolled, such asymmetries confound the interpretation of ChatGPT's contribution to learning outcomes. We do want to highlight that matching instructional features

may often be an approximation rather than a fully attainable standard due to the nature of specific interventions. For example, problem-solving support may look different in a ChatGPT condition than in a more traditional condition involving Google search capabilities. The frequency and type of feedback provided to the ChatGPT condition may also look different than teacher-provided feedback in the classroom. The key for researchers is to describe exactly what occurred in both conditions and to temper claims based on what was controlled and what differed between conditions.

Third, researchers should systematically control and report exposure time and practice opportunities that align with the dependent measure. Unequal time-on-task or differential familiarity with the outcome measure introduces alternative explanations that challenge causal attribution. Standardized learning durations and parallel practice structures are essential to rule out dosage or test-rehearsal effects. However, if researchers are specifically interested in efficiency, they should explicitly justify why time-on-task or dosage will not be controlled.

Fourth, studies should employ direct, validated measures of learning that go beyond face validity or self-report (Martella et al., 2026). Performance assessments, transfer tasks, and structured rubrics can offer more robust insights into cognitive learning outcomes than perceived usefulness or motivational indicators alone.

Fifth, if one were to be interested in whether ChatGPT is more efficient for learning, this variable would need to be the main focus of the study and explicitly mentioned as a difference between conditions. In our sample of studies, time-on-task was often not discussed or isolated as a variable of interest. As such, it becomes a confound that could be problematic for the claims researchers can make about the tool. Future research could involve the investigation of efficiency to determine if one affordance of ChatGPT relates to instructional time.

Sixth, future research should disaggregate the functions that ChatGPT fulfills within an intervention. Rather than treating ChatGPT as a monolithic treatment, it is crucial to examine which specific instructional roles it assumes (e.g., feedback provider, ideation partner, or metacognitive prompt) and to investigate their differential effects. This may require more granular study designs that isolate pedagogical components or employ microgenetic approaches to trace how interactions with ChatGPT shape learning over time.

Finally, researchers may wish to expand on the present review by conducting an additional search of the ChatGPT literature beyond those studies included in the two examined meta-analyses or may wish to analyze the comparability of experimental and control conditions across all AI-related studies in education to further examine the rigor of the literature base.

Taken together, these recommendations aim to raise the methodological standards in the emerging field of generative AI in education and to support a more nuanced, pedagogically grounded understanding of ChatGPT's potential and limitations.

Limitations of the Review

This review is subject to several limitations that should be considered when interpreting its findings. First, our analysis was restricted to the 63 comparisons included in Deng et al.'s (2025) meta-analysis and the 46 comparisons included in Wang and Fan's (2025) meta-analysis. As such, we did not incorporate studies excluded from their synthesis, nor did we consider unpublished or gray literature beyond their own respective search strategies. It is possible that more methodologically rigorous studies exist outside of this corpus, and, if so, their exclusion limits the generalizability of our conclusions. However, given that the studies in Deng et al. (2025) and Wang and Fan (2025) were used to make specific conclusions about ChatGPT's effectiveness, our analysis of these studies provides evidence of the need for more rigorous work

to take place before meta-analyses are conducted. Future reviews should consider broadening their inclusion criteria to encompass gray literature and unpublished studies, which may provide a more comprehensive picture of the current evidence base.

Second, our coding decisions were constrained by the level of detail provided in the primary studies. In many cases, especially regarding instructional methods or time-on-task, information was sparse or ambiguously reported. While we applied a consistent coding framework adapted from Lawson et al. (2024), some classifications relied on interpretive judgment in the absence of clear procedural descriptions. This may have led to conservative coding of the variables. However, we did attempt to reduce this judgment by providing a *difficult to determine* code when information was too limited. Future reviews could contact study authors to clarify ambiguous procedural details or request missing information, thereby improving the accuracy and depth of the methodological assessment.

Third, although we adopted a systematic coding scheme, we must note a limitation concerning the interrater agreement. Agreement levels in early stages revealed the inherent complexity of evaluating equivalence across diverse experimental designs. This complexity was particularly prevalent for Category 4, in which the initial interrater agreement fell below the 80% acceptable threshold. This lower agreement was largely due to the ambiguity in how primary studies reported certain methodological information (e.g., for time on task, comparisons could report maximum possible time given to participants or the actual time of the lesson). The authors went through many rounds of coding and discussion to ensure coding categories were clear and that coding occurred in a consistent fashion. The authors resolved discrepancies through category-specific consensus discussions. For example, in Category 4, we standardized the “likely yes” code for overlapping ranges of time spent on task. Agreement levels within each category

reflected some of the complexities of data extraction within each article. Due to the complexities and concerns around low interrater agreement in initial coding sessions, the review took a rigorous approach by having each article double coded for each coding category; any and all discrepancies were discussed by the research team until resolution was reached. The interpretation of results may be tempered given the complexity of the coding process.

Fourth, our analysis for our second research question focused exclusively on cognitive learning outcomes assessed via direct measures. While this focus aligns with our goal of evaluating claims about ChatGPT's educational effectiveness, we cannot speak to how comparability of conditions relates to the results found for affective, motivational, or behavioral dimensions that are equally relevant for understanding the broader educational impact of generative AI.

Conclusion

Our findings point to two clear and consistent patterns: many studies on ChatGPT in education suffer from (a) missing, pertinent methodological information, particularly in terms of instructional features and time on task, and (b) methodological confounds. As such, the frequently reported performance benefits of ChatGPT must be interpreted with caution. The comparability findings suggest that we must be extremely cautious when making determinations about the impact of ChatGPT itself on learning. At this point, it is difficult to know whether benefits come from the use of ChatGPT or whether the medium facilitates pedagogical practices. The rapid adoption of ChatGPT in educational contexts makes rigorous theory-informed, and transparent research more urgent than ever. Without methodological clarity, the field risks reproducing the same cycles of hype and disillusionment that have accompanied earlier waves of educational technology. Moving forward, research must prioritize not only whether ChatGPT

works but how, for whom, and under what pedagogical conditions it does so. Only through such an approach can we generate findings that are both scientifically credible and educationally actionable.

References

*Included in Deng et al. (2025) and/or Wang and Fan (2025) as well as referenced in manuscript

Anderson, S. C., Humlum, M. K., & Nandrup, A. B. (2016). Increasing instruction time in school does increase learning. *Proceedings of the National Academy of Sciences,* 113, 7481-7484. https://doi.org/10.1073/pnas.1516686113

*Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023a). 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

*Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023b). Better by you, better than me, ChatGPT3 as writing assistance in students' essays. *ArXiv.* https://arxiv.org/abs/2302.04536

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, Ö., and Mariman, R. (2024). Generative AI can harm learning. *Wharton School Research Paper*. https://doi.org/10.2139.ssrn.4895486

Bozkurt, A. (2020). Educational Technology Research Patterns in the Realm of the Digital Knowledge Age. *Journal of Interactive Media in Education, 2020*(1), 18. https://doi.org/10.5334/jime.570

Buchner, J., & Kerres, M. (2023). Media comparison studies dominate comparative research on augmented reality in education. *Computers & Education, 195*, 104711. https://doi.org/10.1016/j.compedu.2022.104711

*Chan, S., Lo, N., & Wong, A. (2024). Generative AI and essay writing: Impacts of automated feedback on revision performance and engagement. *rEFLections, 31*(3), 1249-1284. https://doi.org/10.61508/refl.v31i3.277514

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445–459. https://doi.org/10.3102/00346543053004445

Clark, R. E. (1994). Media Will Never Influence Learning. *Educational Technology Research and Development, 42*(2), 21–29.

*Darmawansah, D., Rachman, D., Febiyani, F., & Hwang, G.-J. (2025). ChatGPT-supported collaborative argumentation: Integrating collaboration script and argument mapping to enhance EFL students' argumentation skills. *Education and Information Technologies, 30*, 3803–3827. https://doi.org/10.1007/s10639-024-12986-4

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, Article 105224. https://doi.org/10.1016/j.compedu.2024.105224

Freeman, S., Eddy, S. L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., & Wenderoth, M.P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410-8415. https://doi.org/10.1073/pnas.1319030111

Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., Järvelä, S., Mavrikis, M., & Rienties, B. (2024). The promise and challenges of generative AI in education. *Behaviour & Information Technology*, 1–27. https://doi.org/10.1080/0144929X.2024.2394886

Godwin, K. E., Seltman, H., Almeda, M., Skerbetz, M. D., Kai, S., Baker, R. S., & Fisher, A. V. (2021). The elusive relationship between time on-task and learning: Not simply an issue of measurement. *Educational Psychology, 41*(4), 502–519. https://doi.org/10.1080/01443410.2021.1894324

Hastings, N. B., & Tracey, M. W. (2005). Does media affect learning: Where are we now? *TechTrends, 49*(2), 28–30. https://doi.org/10.1007/BF02773968

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77-89. https://doi.org/10.1080/19312450709336664

Honebein, P. C., & Reigeluth, C. M. (2021). To prove or improve, that is the question: The resurgence of comparative, confounded research between 2010 and 2019. *Educational Technology Research and Development*, *69*(2), 465–496. https://doi.org/10.1007/s11423-021-09988-1

*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*, *29*, 22969–22990. https://doi.org/10.1007/s10639-024-12754-4

*Jafarian, N. R., & Kramer, A. W. (2025). AI-assisted audio-learning improves academic achievement through motivation and reading engagement. *Computers & Education: Artificial Intelligence, 8*. https://doi.org/10.1016/j.caeai.2024.100357

*Karaman, M. R., & Goksu, I. (2024). Are lesson created by ChatGPT more effective? An experimental study. *International Journal of Technology in Education, 7*(1), 107-127. https://doi.org/10.46328/ijte.607

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T.,

Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103,* 102274. https://doi.org/10.1016/j.lindif.2023.102274

Klahr, D. (2013). What do we mean? On the importance of not abandoning scientific rigor when talking about science education. *Proceedings of the National Academy of Sciences, 110*(3), 14075–14080. https://doi.org/10.1073/pnas.1212738110

Kozma, R. B. (1991). Learning With Media. *Review of Educational Research, 61*(2), 179–211.

Kozma, R. B. (1994). Will media influence learning? Reframing the debate. *Educational Technology Research and Development, 42*(2), 7–19. https://doi.org/10.1007/BF02299087

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of Intelligence Tutoring Systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42-78. https://doi.org/10.3102/0034654315581420

Lawson, A. P., & Martella, A. M. (2023). Critically reflecting on the use of immersive virtual reality in educational settings: What is known and what has yet to be shown? *Journal of Applied Learning & Teaching, 6*(2), 1–13. https://doi.org/10.37074/jalt.2023.6.2.35

Lawson, A. P., Martella, A. M., LaBonte, K., Delgado, C. Y., Zhao, F., Mayer, R. E., Gluck, A. J., Munns, M. E., & Wells, A. (2024). Confounded or controlled? A systematic review of media comparison studies involving immersive virtual reality for STEM education. *Educational Psychology Review, 36,* Article 69. https://doi.org/10.1007/s10648-024-09908-8

Lehmann, M., Cornelius, P. B., & Sting, F. J. (2024). AI meets the classroom: When does ChatGPT harm learning?. *Available at SSRN 4941259*.

*Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing. *Assessment and Evaluation in Higher Education, 49*(5), 616-633. https://doi.org/10.1080/02602938.2024.2301722

*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

Martella, A. M., Lawson, A. P., Martella, R. C., LaBonte, K., Zhao, F., Delgado, C. Y., Wells, A., Gluck, J. A., Munns, M. E., & Mayer, R. E. (2026). Systematically reviewing the rigour of immersive virtual reality research in STEM education: A deep dive into threats to internal validity. *Journal of Computer Assisted Learning*, *42*(1), e70165. https://doi.org/10.1002/jcal.70165

Martella, A. M., Martella, R. C., Yatcilla, J. K., Newson, A., Shannon, E. N., & Voorhis, C. (2023). How rigorous is active learning research in STEM education? An examination of key internal validity controls in intervention studies. *Educational Psychology Review*, *35*(4), Article 107. https://doi.org/10.1007/s10648-023-09826-1

Martella, R. C., Nelson, J. R., Morgan, R. L., & Marchand-Martella, N. E. (2013). *Understanding and interpreting educational research.* The Guilford Press.

Martella, A. M., & Schneider, D. W. (2024). A reflection on the current state of active learning research. *Journal of the Scholarship of Teaching and Learning*, *24*(3), 119–136. https://doi.org/10.14434/josotl.v24i3.35263

Mason, E. N., & Smith, R. A. (2020). Tracking intervention dosage to inform instructional decision making. *Intervention in School and Clinic, 56*(2), 92–98. https://doi.org/10.1177/1053451220914897

Mayer, R. E. (2019). Computer Games in Education. *Annual Review of Psychology*, *70*, 531–549. https://doi.org/doi.org/10.1146/annurev-psych-010418- 102744

Mayer, R. E. (2020). *Multimedia Learning* (Third Edition). Cambridge University Press. cambridge.org/9781107187504

Mishra, P., Koehler, M. J., & Kereluik, K. (2009). The Song Remains the Same: Looking Back to the Future of Educational Technology. *TechTrends*, *53*(5), 48–53. https://doi.org/10.1007/s11528-009-0325-3

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

*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

Ramsey, J. L., & West, R. E. (2023). A Recent History of Learning Design and Technology. *TechTrends*, *67*(5), 781–791. https://doi.org/10.1007/s11528-023-00883-5

Reeves, T. C., & Oh, E. G. (2017). The goals and methods of educational technology research over a quarter century (1989–2014). *Educational Technology Research and Developmen*t, 65, 325-339. https://doi.org/10.1007/s11423-016-9474-1

Reich, J. (2020). *Failure to Disrupt: Why Technology Alone Can't Transform Education*. Harvard University Press.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Karpicke, J. D. (2018). Reflections on the resurgence of interest in the testing effect. *Perspectives on Psychological Science, 13*(2), 236–241. https://doi.org/10.1177/1745691617718873

*Roganović, J. (2024). Familiarity with ChatGPT features modified expectations and learning outcomes of dental students. *International Dental Journal, 76*(6), 1454-1462. https://doi.org/10.1016/j.identj.2024.04.012

*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*, 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

*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. In Cheng, YP., Pedaste, M., Bardone, E., & Huang, YM. (Eds.), *Innovative technologies and learning* (pp. 205–214). Springer. https://doi.org/10.1007/978-3-031-65884-6_21

Surry, D. W., & Ensminger, D. (2001). What's Wrong with Media Comparison Studies? *Educational Technology*, *41*(4), 32–35. https://www.jstor.org/stable/44428679

Vanlehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31*(1), 3–62. https://doi.org/10.1080/03640210709336984

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 12*. https://doi.org/10.1057/s41599-025-04787-y

Wang, Y., Liu, W., Yu, X., Li, B., & Wang, Q. (2024). The impact of virtual technology on students' creativity: A meta-analysis. *Computers & Education, 215*, 105044. https://doi.org/10.1016/j.compedu.2024.105044

Wedel, K. (2021). Instruction time and student achievement: The moderating role of teacher qualifications. *Economics of Education Review, 85*, 102183. https://doi.org/10.1016/j.econedurev.2021.102183

Weidlich, J., Gašević, D., Drachsler, H. & Kirschner, P. (2025). ChatGPT in education: An effect in search of a cause. *Journal of Computer Assisted Learning, 41*, e70105. https://doi.org/10.1111/jcal.70105

*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,* 47(3), 445–449. https://doi.org/10.1080/0142159X.2024.2339412

*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, Italy. https://doi.org/10.1145/3664934.3664946

Yan, L, Grieff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour, 8*(10), 1829-1850. https://doi.org/10.1038/s41562-024-02004-5

Yan, L., Martinez-Maldonado, R., Jin, Y., Echeverria, V., Milesi, M., Fan, J., ... & Gašević, D. (2025). The effects of generative AI agents and scaffolding on enhancing students' comprehension of visual learning analytics. *Computers & Education*, 105322.

Zhai, X. (2023). ChatGPT and AI: The game changer for education. *Shanghai Education,* 16-17. https://ssrn.com/abstract=4389098

*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. The 6th International Conference on Computer Science and Technologies in Education, Xi'an, China. https://doi.org/10.1109/CSTE62025.2024.00035

Table 1

Results of Assessment Criteria for Comparisons Between ChatGPT and Control Conditions

Coding Categories	Deng et al. (2025) N = 63 (unless otherwise noted)	Wang and Fan (2025) N = 46 (unless otherwise noted)
Category 1: Operational Definitions		
Across Both Conditions		
At least partial in both conditions	43 (68.25%)	32 (69.57%)
No in at least one condition	20 (31.75%)	14 (30.43%)
ChatGPT Condition Only		
Yes	32 (50.79%)	25 (54.35%)
Partial	22 (34.92%)	14 (30.43%)
No	9 (14.29%)	7 (15.22%)
Control Condition Only		
Yes	24 (38.10%)	21 (45.65%)
Partial	19 (30.16%)	11 (23.91%)
No	20 (31.75%)	14 (30.43%)
Category 2: Match Instructional Methods		
Yes	11 (17.46%)	4 (8.70%)
No	19 (30.16%)	17 (36.96%)
Difficult to Determine	33 (52.38%)	25 (54.35%)
Category 3: Matched Practice with Dependent Measure		
Not Applicable (no direct learning measure present)	15 (23.81% of 63)	10 (21.74% of 46)
Applicable (direct learning measure present)	48 (76.19% of 63)	36 (78.26% of 46)
Yes	14 (29.17% of 48)	15 (41.67% of 36)
No	6 (12.50% of 48)	4 (11.11% of 36)
Difficult to Determine	28 (58.33% of 48)	17 (47.22% of 36)

Coding Categories	Deng et al. (2025) N = 63 (unless otherwise noted)	Wang and Fan (2025) N = 46 (unless otherwise noted)
Category 4: Matched Time Spent Learning the Content		
Yes	19 (30.16%)	13 (28.26%)
Likely Yes	2 (3.18%)	3 (6.52%)
No	2 (3.18%)	1 (2.17%)
Likely No	2 (3.18%)	1 (2.17%)
Difficult to Determine	38 (60.32%)	28 (60.87%)
Category 5: Direct Measure of Learning		
Yes	41 (65.08%)	28 (60.87%)
No	14 (22.22%)	10 (21.74%)
No because ChatGPT was used during posttest	7 (11.11%)	8 (17.39%)
Difficult to Determine	1 (1.59%)	0 (0.00%)
Category 6: Results of the Study for Direct Measures of Learning	*(n = 41)*	*(n = 28)*
ChatGPT > Conventional	27 (65.85%)	18 (64.29%)
Conventional < ChatGPT	2 (4.88%)	1 (3.57%)
Tied	9 (21.95%)	6 (21.43%)
Mixed	1 (2.44%)	2 (7.14%)
Inconclusive	2 (4.88%)	1 (3.57%)

Figure 1

Percentage of comparisons that violated instructional-feature controls when no benefit of the doubt was given

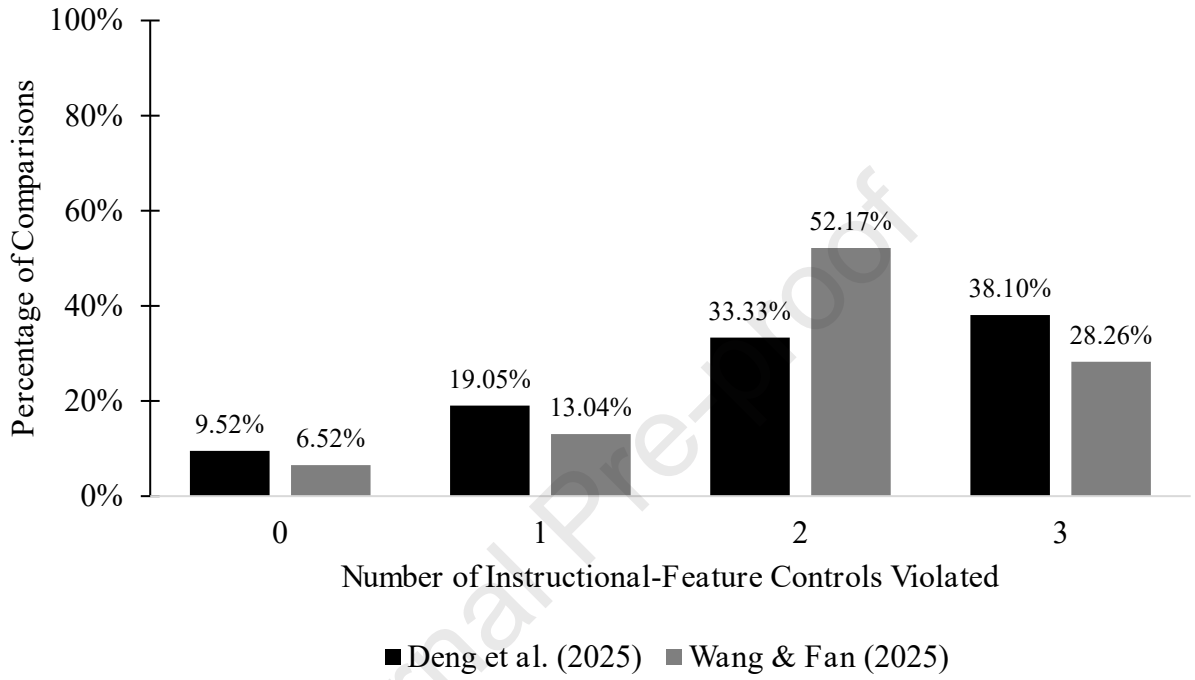


Figure 2

Percentage of comparisons that violated instructional-feature controls when benefit of the doubt was given

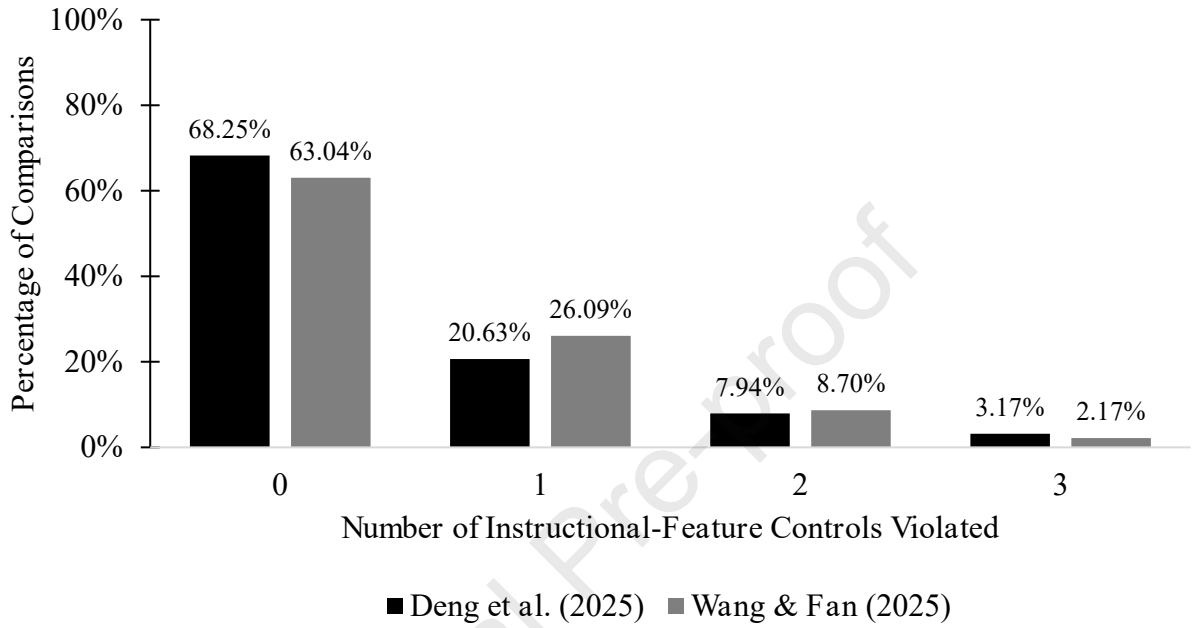


Figure 3

Counts of ChatGPT and Control Conditions Deemed Comparable, Not Comparable, and Difficult to Determine Across Articles (n = 41) in Deng et al. (2025)

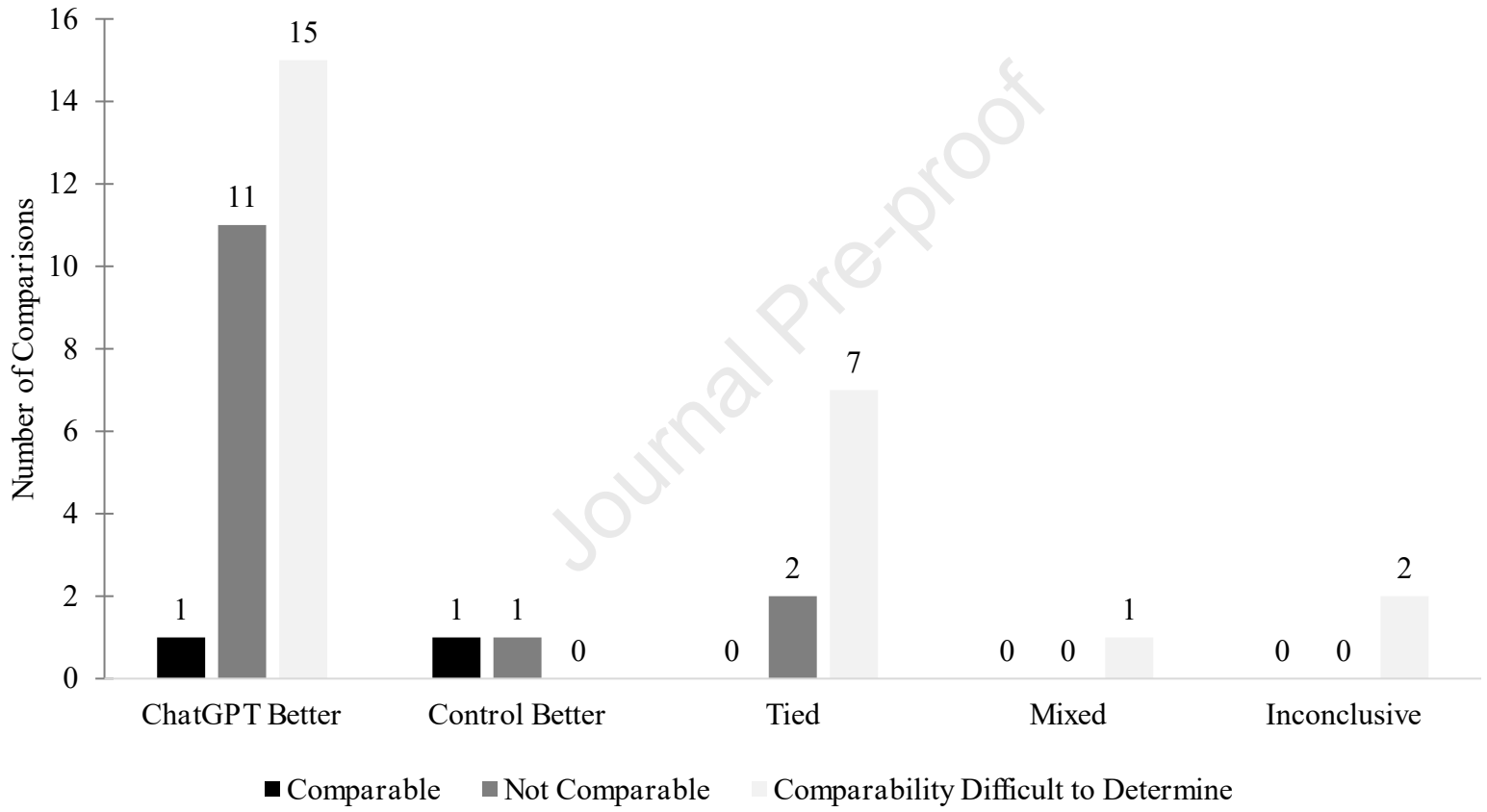
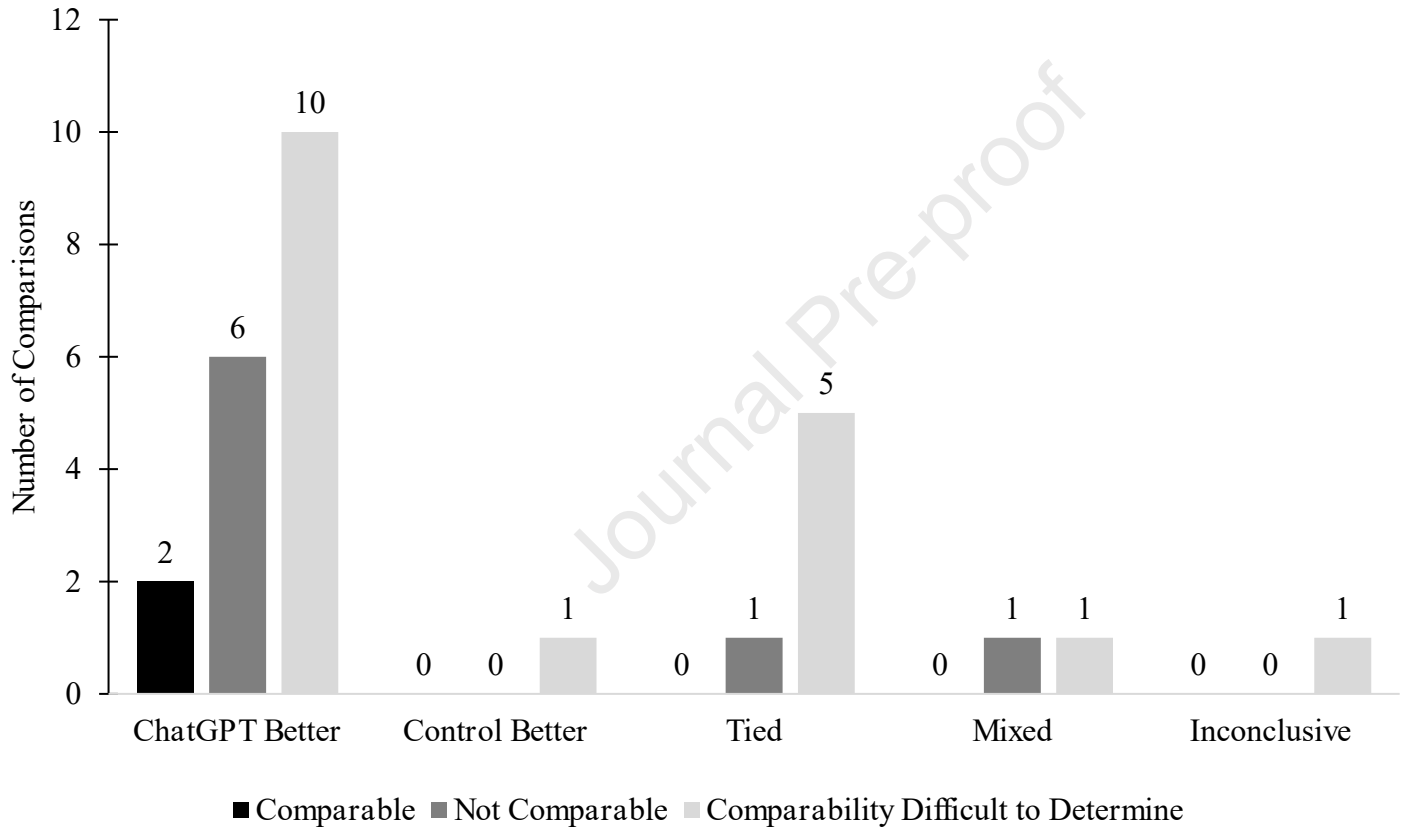


Figure 4

Counts of ChatGPT and Control Conditions Deemed Comparable, Not Comparable, and Difficult to Determine Across Articles (n = 28) Wang and Fan (2025)



Highlights

Two meta-analyses on media comparison research investigating the impact of ChatGPT on learning were critically analyzed.
Many media comparisons are not controlled in time on task, instructional methods, and practice with the dependent measure of learning across conditions.
Media comparisons that find ChatGPT to be effective for learning tend to be confounded.