Journal of Computer Assisted Learning

WILEY

Journal of Computer Assisted Learning

ORIGINAL ARTICLE OPEN ACCESS

ChatGPT in Education: An Effect in Search of a Cause

J. Weidlich[1,2] | D. Gašević[3] | H. Drachsler[4,5,6] | P. Kirschner[6,7]

[1]University of Zurich, Zurich, Switzerland | [2]Zurich University of Teacher Education, Zurich, Switzerland | [3]Monash University, Clayton, Victoria, Australia | [4]DIPF—Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany | [5]Goethe Universität, Frankfurt am Main, Germany | [6]Open University of the Netherlands, Heerlen, the Netherlands | [7]Thomas More University of Applied Sciences, Antwerpen, Belgium

Correspondence: J. Weidlich (joshua.weidlich@ife.uzh.ch)

Received: 25 March 2025 | **Revised:** 17 June 2025 | **Accepted:** 3 August 2025

ABSTRACT

Background: As researchers rush to investigate the potential of AI tools like ChatGPT to enhance learning, well-documented pitfalls threaten the validity of this emerging research. Issues of media comparison research, where the confounding of instructional methods and technological affordances is unrecognised, may render effects uninterpretable.

Objectives: Using a recent meta-analysis by Deng et al. (*Computers & Education*, 227, 105224) as an example, we revisit key insights from the media/methods debate to highlight recurring conceptual challenges in ChatGPT efficacy studies.

Methods: This conceptual article contrasts nascent ChatGPT research with the more established literature on Intelligent Tutoring Systems to identify three non-negotiable considerations for interpretable effects: (1) descriptions of the precise nature of the experimental treatment and (2) the activities of the control group, as well as (3) outcome measures as valid indicators of learning. To provide some initial evidence, we audited a subset of primary experiments included in Deng et al.'s meta-analysis, demonstrating that only a small minority of studies satisfied all three non-negotiable considerations.

Results and Conclusions: Loosely defined treatments, mismatched or opaque controls, and outcome measures with unclear links to durable learning obscure causal claims of this emerging literature. Observed gains cannot, at this time, be confidently attributed to ChatGPT, and meta-analytics effect sizes may over- or understate its benefits. Progress, we argue, will require rigorous designs, transparent reporting, and a critical stance toward "fast science."

1 | Introduction

With the rapid proliferation of powerful generative artificial intelligence (AI) systems like ChatGPT, the potential of these technologies for education has become the focus of intensive research efforts (Giannakos et al. 2024; Yan et al. 2025). With every update, these systems increasingly evoke the holy grail of highly adaptive, personalised educational agents. Understandably, researchers and educators seek to assess both the promises and limitations of these tools for learning and teaching. However, the rush to generate findings risks premature conclusions based on weak evidence. Additionally, long-standing methodological and conceptual pitfalls—well documented in decades of media comparison research—threaten to obscure a clear picture of how generative AI can enhance learning. Addressing these challenges early on is necessary to build a robust and meaningful evidence base for the use of AI in education.

A recently published meta-analysis by Deng et al. (2025) exemplifies the urgency of many researchers to synthesise research on ChatGPT within 2 years of its launch. Examining this early synthesis raises concerns about the quality of the primary studies included in the analysis. While early studies can provide valuable exploratory insights, they often lack the methodological features necessary for drawing robust causal conclusions. This is particularly concerning in the case of Deng et al. (2025), as both the meta-analysis and many of its included studies explicitly make causal claims, despite methodological limitations. Indeed, the authors of the meta-analysis themselves acknowledge that many of these studies lack essential features such as

Summary

- What is currently known about this topic?
  - Early studies report large learning gains from using ChatGPT in education.
  - Many studies appear to compare unclear ChatGPT treatments with unknown controls.
  - Media comparison pitfalls render effects uninterpretable.
- What does this paper add?
  - Paper identifies three non-negotiables: well-defined treatment and control, and valid learning measures.
  - Rapid audit of a subset of studies finds few studies meet all three.
  - Contrasts ChatGPT studies with more established Intelligent Tutoring Systems work.
  - Warns of "fast science" and calls for reflective, more deliberate research practices.
- Implications for practice/or policy
  - Claims of ChatGPT's effectiveness for learning are currently unfounded.
  - Educators should carefully consider the specific instructional features of their planned ChatGPT intervention.
  - A more mature research literature is needed before recommendations for educational practice are warranted.

power analyses, random assignment, and pre-test measures to assess knowledge change over time.

However, individual study flaws are not the subject of this article. Instead, we examine three broader *conceptual* considerations: the contents of the experimental treatment, the nature of the control group, and the dependent variable as a reflection of student learning. While revising this manuscript, a second meta-analysis on ChatGPT in education was published. Wang and Fan (2025) analysed studies from a similar period and reported comparable findings to those of Deng and colleagues: a substantial positive effect of ChatGPT on learning. Although our discussion centres on Deng et al. (2025) as a key example, the concerns we raise apply equally to Wang and Fan (2025) and future research aiming to assess the effects of ChatGPT on learning. To gauge how widespread these issues are, we drew on a subset of studies from Lawson et al. (forthcoming), taking a closer look at the primary studies included in Deng et al.'s meta-analysis. Our analysis focused on how clearly the treatment and control conditions were described and what kind of learning outcome measures were used.

As a point of comparison, the more established literature on Intelligent Tutoring Systems (ITS) also utilises AI systems in education and does this successfully by leveraging decades of instructional design research to embed AI within robust pedagogical frameworks. Meta-analyses of ITS studies typically yield interpretable findings due to clarity regarding the aforementioned conceptual considerations (Kulik and Fletcher 2016; Ma et al. 2014; Steenbergen-Hu and Cooper 2014). Comparing these two research traditions helps highlight both the strengths of established AI applications in education and the risks emerging in newer lines of inquiry.[1]

Generative AI has the potential to transform education, but requires a deliberate, measured research approach to establish conditions of success or failure. Considering the importance of "getting it right," this article aims to support researchers in conducting research that is fit for moving our understanding forward. We begin by revisiting key insights from the media/methods debate that remain relevant today. We then compare two distinct research lines, identifying strengths in the more established literature and exposing methodological pitfalls in the emerging generative AI research. While these pitfalls affect both primary and secondary research, we focus on two recent meta-analyses to extract three key considerations for conducting interpretable and meaningful studies in this space. We conclude with reflections on the risks of fast science and the necessity of methodological rigour in AI education research.

2 | The Problem of Confounding in Media Comparisons

Richard E. Clark's seminal 1983 article, *Reconsidering Research on Learning from Media*, critically examines the relationship between media and learning, challenging the assumptions underlying much of the then-existing research in this area. With it, Clark (1983) kickstarted a cycle of papers known as the Media/Methods debate, including Robert Kozma (1991, 1994a, 1994b), the main opponent of Clark's position (Clark 1983, 1994a, 1994b). The debate culminated in a 1994 special issue of *Educational Technology Research and Development*, where several other researchers weighed in Jonassen et al. (1994), Morrison (1994), Reiser (1994), Ross (1994), Shrock (1994), and Tennyson (1994).[2]

In short, Clark (1983) argued that media function solely as vehicles for delivering instruction and do not, in themselves, influence student achievement, famously likening media to trucks delivering groceries: while the truck facilitates the delivery, it does not alter the nutritional value of the groceries. Similarly, the efficacy of learning is determined by the instructional method, not the medium through which it is delivered. While he allowed for efficiency gains by choosing an appropriate medium for a particular learning goal, he remained firm that media are, in principle, replaceable by another medium with similar learning gains (Clark 1994a, 1994b). Kozma (Kozma 1991; Kozma 1994a; Kozma 1994b), on the other hand, rejects the notion that media are neutral carriers, arguing instead that their unique affordances intertwine with instructional methods and learners' cognition.

Although Clark and Kozma each capture a valuable—and ultimately complementary—challenge of our field, Clark's stance feels more urgent in the current landscape of ChatGPT media comparison studies.[3] When researchers run media comparison experiments (rather than value-added or aptitude-treatment-interaction designs; Buchner and Kerres 2023), they implicitly assume that the technology can be inserted into—or removed from—a learning setting without altering the underlying pedagogy (Honebein and Reigeluth 2021). In other words, researchers are voting with their feet—or, more precisely, with their researchdesigns—revealingg an assumption that medium and method are separable, in contrast to Kozma's (1991) arguments for their interdependence.

Once that assumption is in play, Clark's methodological lesson follows: if researchers wish to draw any causal conclusions from these designs, they must aim to keep the instructional method constant and scrutinise what, if anything, the medium itself contributes. However, new media or technologies are often introduced alongside new instructional methods, interface designs, or activity structures—what Eronen (2020) calls "fat-handed" interventions. Such studies fail to surgically remove or add the variable of interest without altering the broader instructional mix, rendering their results unfit for interpretation, at least in the way intended by their authors.[4] A similar confusion can arise in program-based studies, as Zhang et al. (2022) illustrate, where learning gains are credited to an inquiry-based design. However, such programs bundle many components, each potentially contributing to the effect.

Tennyson (1994), summarising the debate, provides the analogy of a big wrench, where naïve media comparisons imply the currently hot-topic medium as a panacea solution for learning. In reality, however, the big wrench view obscures what is actually driving learning, as potentially efficacious features of the treatment combined with the medium remain unexamined. While the media at that time was instructional television or early computer-based instruction, it is easy to find more contemporary examples. The issue of confounding was found to plague evaluations of the effectiveness of distance education (Bernard et al. 2009; Clark 2000; Lou et al. 2006) and, more recently, Augmented Reality and Virtual Reality. Buchner and Kerres (2023) conclude from their systematic review that 80% of studies on the effects of Augmented Reality for learning are media comparisons. Similarly, Lawson et al. (2024) conducted an in-depth analysis of media comparison studies on Virtual Reality in STEM education, finding that only 26% of studies controlled for confounding across multiple criteria.

3 | Conditions for Interpretable Meta-Analytic Effects Using the Example of Intelligent Tutoring Systems

There is a well-established tradition of research on AI-based systems explicitly designed to foster learning. These systems, called Intelligent Tutoring Systems (ITS), integrate AI into well-defined pedagogical models tailored to domain-specific learning objectives. For example, Cognitive Tutors are a kind of ITS, often based on Anderson's ACT-R theory (Anderson 1982; Anderson et al. 1995) and provide step-by-step guidance as learners acquire and develop problem-solving skills through practice (Koedinger and Corbett 2006). A key component of most ITS is a learner model, which changes throughout the learning trajectory as students make errors or acquire proficiency. Other key components of their architecture are domain models to represent the content and pedagogical models to make instructional moves based on student behaviour.

Through these features, ITS offer individualised instruction, immediate feedback, and adaptivity, making them powerful tools for enhancing specific cognitive and metacognitive skills. Meta-analyses consistently show that ITS can yield substantial learning gains, although they remain slightly outperformed by human tutoring (VanLehn 2011). For example, the meta-analyses by Steenbergen-Hu and Cooper (2014), Ma et al. (2014), and Kulik and Fletcher (2016) demonstrate that ITS outperform traditional instruction, with effect sizes often ranging between 0.32 and 0.66. Using Kulik and Fletcher (2016) as an example, we outline the hallmarks of a compelling meta-analytic synthesis of a literature focused on identifying the learning effects of a technology.

Note that meta-analytical syntheses are governed by established quality criteria (e.g., Borenstein et al. 2009; Pigott and Polanin 2020) and reporting standards (PRISMA, Moher et al. 2015). As abundant material on criteria and standards is available, our goal is not to contribute to this literature. Instead, we focus here on more fundamental considerations that must be in place before such frameworks can operate.

First and foremost, Kulik and Fletcher (2016) code *the contents of the treatment* groups by including the ITS (e.g., Geometry Tutor, Cognitive Tutor, and ACT Programming Tutor). As these systems are described in the primary studies, the instructional features of the treatment are entirely recoverable for careful readers. Additional information, like the duration of the treatment and whether the ITS used step-based tutoring or substep-based tutoring, allows for a relatively holistic picture of what students experienced in the treatment.[5] Importantly, while factors like the student population, teaching subject, and other aspects varied, ITS shared several features that make these treatments similar and, thus, identifiable as one treatment unit. As mentioned, these features include individualised instruction, immediate feedback, and adaptivity.

Further, Kulik and Fletcher (2016) attend to *the nature of the control group*. Most of the studies in their analysis had a conventional control group (i.e., students in the control condition received conventional instruction or instruction closely approximating conventional teaching). Studies with a no-treatment control group, human tutor control group, or other designs were excluded from the primary analyses. These studies received additional coding and were analysed separately, given that significant differences in a control condition fundamentally change the nature of the comparison.

This clear and transparent delineation of treatment and control groups allows for a meaningful interpretation of effect sizes. In Kulik and Fletcher (2016), ITS, with their integrated instructional features, yielded an effect size of 0.66 standard deviations in learning outcomes compared to conventional teaching. Given that conventional teaching, by definition, approximates many learning scenarios in educational practice, this effect size indicates a substantial improvement over the status quo. Based on it, one could also estimate hypothetical achievement gains in grades over time should ITS be employed on a larger scale. Alternatively, the effect size could be translated into metrics like learning gains expressed in years of schooling or percentiles (Baird and Pane 2019). Crucially, these translations only make sense when the control group plausibly maps onto realistic educational settings.

For ITS, thus, the active ingredient is not a single variable but an integrated set of instructional features. The question of which feature caused the learning effects remains unanswered, but this is only problematic if the goal were to assess only a single

variable instead. In ITS research, this does not hinder interpretation, as real-world use also involves applying these features as a package. We already know that individual principles (e.g., immediate feedback) foster learning because this research base underpins ITS design (see Anderson et al. 1995; Corbett et al. 1997; Graesser et al. 2018). The key question is whether—and to what extent—automated systems with these integrated properties support learning, not whether individual components work in isolation. Thus, the focus shifts from testing tenets of instructional theories, which guided ITS design, to confirming their efficacy in practice.

This brings us to an essential aspect of what makes an experimental comparison meaningful. The comparison should be substantively interesting, given the underlying research aims. For example, the question "Is Treatment X more effective than the status quo Y?" requires a business-as-usual control group in which activities closely mirror classroom realities. If the underlying research aim is different (i.e., to confirm a particular theoretical tenet or to establish a baseline), then a meaningful comparison may require a different approach. For example, a passive control group (where students receive no treatment) may be meaningful in basic cognitive research to establish a baseline for forgetting in retrieval practice effects (Karpicke and Roediger 2008). However, for most research in educational technology and related fields, an entirely passive control group will not yield meaningful comparisons.

A final hallmark of establishing the effectiveness of ITS, or of any other educational innovation, is *a clear conception of the dependent variable, that is, the measure of learning*. Kulik and Fletcher (2016) distinguish between locally developed posttests and standardised tests. Considering again what is substantively meaningful, it follows that a standardised test is a more informative source of information regarding the potential effects of ITS in many real-world education contexts. As locally developed posttests are more intertwined with features of a specific learning context, their findings cannot be expected to generalise to the same extent as findings from standardised tests, which, by definition, are relevant to all students of the population. Crucially, to establish an effect on *learning*, as opposed to task performance, the measurements must occur after the fact, and must be identical for both treatment and control, and, ideally, should be standardised and not experimenter generated. If measured during treatment, differences likely reflect situational (dis)advantages due to the (non)availability of the treatment rather than durable learning. Non-identical measures between groups render results wholly uninterpretable.

To summarise, we have identified three basic considerations for interpretable effects: (1) a clear conception of the treatment, (2) the control group, and (3) a dependent variable that validly represents learning. But what justifies these three considerations in particular? We argue that they form the most basic, non-negotiable elements to claim the effectiveness of any given technology for learning.

The importance of a clear description of the treatment traces back to construct validity (Shadish et al. 2002); an insufficiently defined treatment construct does not lend itself to valid inferences about its effects. An opaque treatment invites an array of alternative causal influences and, thus, causal explanations alongside the treatment of interest, leading to tenuous inferences. Or, simply put, if we do not know how ChatGPT was used—or what else was happening—we cannot meaningfully interpret the results.

Equally essential is a well-defined control group, which operationalizes the counterfactual (Hernán and Robins 2020). In our case, that is the learning that would occur had the focal technology been removed without fundamentally altering the learning experience (Salomon 1991). This notion is captured by the *ceteris paribus* assumption ("all other things being equal") of scientific inquiry (Marshall 1890; Heckman and Pinto 2021), which reminds us that meaningful comparisons depend on holding other factors constant. A well-chosen control group makes a comparison meaningful, whereas unclear or unsuitable control activities may render the study uninterpretable. Thus, deeply thinking about, carefully operationalising, and transparently reporting about the duo of treatment and control is simply a precondition for interpreting causal effects.[6]

Finally, the third consideration simply speaks to the aim of drawing causal conclusions about *learning*, rather than other outcomes such as students' perceptions of learning (see Deslauriers et al. 2019). Without this distinction, we risk mistaking satisfaction or engagement for actual learning gains.

4 | Directions for Meta-Analyses of Research on Learning With ChatGPT

Given the impressive capabilities of current AI systems like ChatGPT, researchers are rallying around the topic to explore their potential role for learning. This fervour is illustrated by a meta-analysis of Deng et al. (2025), who, through their systematic review, identified more than a thousand candidate studies for inclusion. Even more astounding is that they identified 22 already published review articles on ChatGPT in education, prompting the question of the validity of the underlying study corpus and the utility of such early syntheses, considering that ChatGPT, as most users know it, appeared only in November of 2022. Aside from this problem of rushed research, which we will return to briefly at the end, the focus of this section is to show how this emerging literature is prone to media comparison pitfalls.

When attempting to investigate the effects of something like ChatGPT on learning, a perennial stumbling block is that ChatGPT, like other "vanilla" generative AI systems, is a tool and not a method for learning or teaching. Like the truck that cannot determine the nutritional value of its content, ChatGPT cannot determine whether students learn. In the same way that we do not think to ask whether Google Scholar fosters internet literacy or a calculator enhances algebra skills, speaking of the learning effects of ChatGPT amounts to a non sequitur; learning does not per se follow from a tool with an entirely different purpose. To support learning, ChatGPT, or any AI system for that matter, must be embedded, integrated, and combined with an instructional method, a pedagogy. Understanding this point is paramount for avoiding the pitfalls plaguing this emerging research literature.

Putting aside this issue for a moment, consider that the impressive capabilities of ChatGPT essentially invite cognitive offloading of the very processes usually intended to support learning. For example, in academic writing, students using ChatGPT can now produce passable texts without acquiring the skills needed for adequate performance. Indeed, some early research indicates that precisely through the abilities that make it so impressive, ChatGPT can harm the depth of processing (Stadler et al. 2024) and invite learners to rely on rather than learn from it (Darvishi et al. 2024; Fan et al. 2024). In this vein, Salomon (1990) speaks of the difference between effects *with* and effects *of* a medium. Some early evidence confirms that how ChatGPT was used is the main determinant of its effects. For example, Lehmann et al. (2024) found that students learned when using Large Language Models (LLMs) as personal tutors by engaging in productive back-and-forth. Conversely, learning outcomes were negative when students used the system to shortcut mental processes intended for learning. Similarly, Bastani et al. (2024) reported detrimental effects on learning when ChatGPT was used as a crutch, whereas Yan et al. (2025) found durable learning gains when the AI agent provided scaffolding. While still early, these findings already paint the picture that, as with all other tools (or media, per Clark), the concrete instantiation of how learners can interact with, rely on, and derive insights from ChatGPT varies dramatically.

Given this, it is surprising that Deng et al. (2025), based on 18 primary studies evaluating academic performance, concluded that ChatGPT improves learning. With an effect size of $g = 0.7$, ChatGPT would thus be more impactful than ITS (see, e.g., ES = 0.66 in Kulik and Fletcher 2016), although ITS are designed with the sole purpose of being as effective for learning as possible, while ChatGPT was designed for entirely different reasons. Given this, the authors would need to provide sufficient information in their coding of primary studies to evaluate how these effects were achieved. For example, did the researchers in the primary studies provide ChatGPT as a personalised tutor, as suggested by Lehmann et al. (2024)? But, as outlined in the previous section, additional considerations are required: How exactly did the control condition differ from the treatments? And how reflective of learning are the outcome measures?

The following subsections discuss each consideration in more detail and relate them to the literature synthesised in Deng et al. (2025).

4.1 | What Is the Treatment?

Deng et al.'s (2025) approach to selecting admissible treatment designs would require one of two things for a compelling claim of learning effectiveness. First, stringent inclusion criteria could be used (e.g., only studies where ChatGPT served as a personalised tutor). This criterion should be circumscribed to a degree that allows the authors to make binary inclusion or exclusion decisions. Presumably, this would drastically reduce the corpus of studies available for analysis, which speaks to the benefits of letting a sufficiently rich literature accumulate before aggregating it. From inspecting a sample of the included primary studies, we can determine that this approach was not chosen, as the experimental treatments cover designs as varied as, for example, (a) ChatGPT as a tool for rehearsal of terminology (Hsu 2024), (b) ChatGPT as mainly a feedback source to support self-assessment in writing (Mahapatra 2024), or (c) ChatGPT as an all-round assistant with no restrictions on its usage to support writing (Niloy et al. 2024).

A second approach would be to have lenient inclusion criteria, allowing for various designs that include using ChatGPT in numerous ways and to different degrees, as did Deng et al. (2025). In this scenario, however, it falls upon the meta-analysts to carefully code those features of the treatment that may plausibly affect learning, for example, the prescribed mode of interacting with ChatGPT, the instructional principles enhanced by the usage of ChatGPT, whether the task was ill-structured or highly structured, whether limitations were placed on the usage of ChatGPT, how much relevant domain knowledge was fed into ChatGPT beforehand, and many others. With these codes, an eclectic mix of primary studies could be palatable because these instructional features (i.e., instantiations of the ChatGPT-method combinations) could be included as moderators in the analyses, given sufficient cell sizes. Therefore, the relative efficacy of these combinations could be demarcated, albeit with limitations.[7] While Deng et al. (2025) code some features (e.g., intervention setting, duration, and whether a learning application was ChatGPT-supported or ChatGPT was itself the main feature), these do not even capture the instructional diversity of the treatments by Hsu (2024), Mahapatra (2024), and Niloy et al. (2024), not to speak of the remaining corpus. Without these hallmarks—highly selective inclusion of studies or detailed codes—the treatment in question amounts to a "secret sauce" with unknown ingredients, rendering the effect uninterpretable.

4.2 | What Is the Comparison?

The control condition provides the contrast from which an experimental effect can stand out in sharp relief. In Deng et al. (2025), the inclusion criterion for the control condition is defined solely by the absence of ChatGPT ("The study included at least one control group that did not use ChatGPT [...], table 3, 7). This leniency invites control conditions as varied as no-treatment, one-to-one tutoring with a human, self-assessments, playing video games, and many others. Judging from the overall aim of the meta-analysis, the most appropriate control group would mirror the treatment group, but without ChatGPT. Only then would the contrast yield the distinct shape of ChatGPT, as opposed to a fuzzier, confounded picture. As we know from Clark's (1983) key argument against media comparisons, this is not always trivial, as we often cannot surgically remove ChatGPT without changing other elements.

Granting this, the authors may decide to define strict criteria for exclusion (e.g., the control condition should replicate the instructional features of the treatment to a reasonable extent), as was done by Kulik and Fletcher (2016). In this case, a careful description of the replacement for ChatGPT would be needed. For example, is it replaced by the teacher or the peers, or is it a non-LLM-based AI, or is there no replacement? Again, as an alternative to the rigorous selection of mostly homogenous control groups in their primary studies, the authors may instead opt for

detailed coding to characterise the extent to which the control group design matches the treatment group in terms of instructional features, the business-as-usual learning experiences, and other pedagogically meaningful aspects like teacher or peer involvement as a replacement for ChatGPT. Given a homogenous corpus, an overall meta-analysed effect size would be interpretable on its own. In contrast, in the latter case, the moderator analyses would provide the necessary differentiation to derive insights, depending on whether a sufficiently rich literature has accrued until then.

4.3 | Is This Learning?

The final piece of this puzzle to make this synthesis interpretable is the dependent variable, which needs a set of properties to qualify. The most important property is that it is appropriate for the research question it addresses. If the aim is to identify effects on student learning, an array of potential variables is off the table. For example, student and/or teacher perceptions of learning or self-reports of skills will not do. Learning is commonly defined as the process of acquiring knowledge, skills, attitudes, and behaviours through experience, study, or teaching (Sternberg and Sternberg 2017). A crucial second part of this definition is that the result of this learning process is a relatively enduring change in an individual's understanding, competence, or behaviour (e.g., Schunk 2020). Therefore, self-ratings of learning or abilities are insufficient to demonstrate learning due to the possibility of giving higher self-ratings without an underlying and relatively enduring change in the learner (see e.g., Carpenter et al. 2020). Similarly, some measures of performance will not do if they do not establish that enhanced performance is solely due to learning. In other words, a bona fide learning measure cannot be administered while a student is still experiencing the treatment, as student performance then may either be due to the aforementioned relatively enduring changes (i.e., learning) or due to additional information, resources, or support still available through the treatment. That is, the measurement should be temporally and operationally removed. As we will show, Deng et al. (2025), despite the titular focus on student learning, measure various outcomes, most of which are not learning.[8]

To distinguish categories of outcomes conceptually and analytically, the authors calculated separate models (visually summarised as forest plots in Appendix B of Deng et al. 2025). Given the research aim of establishing learning effects, we focus on the models most closely related to learning, that is, academic performance and higher order thinking propensities.

The first study on academic performance is Ahmed Moneus and Al-Wasy (2024), reporting a whopping effect size of Hedge's $g = 3.1$. As such, this study contributes significantly to the overall academic performance outcome. The study examines how ChatGPT impacts Saudi translators, comparing the quality of translations produced with and without ChatGPT. Importantly, the outcome measure is not translation quality *after* using ChatGPT, but the quality of the text produced *during* human-ChatGPT collaboration. Thus, the case is clear; the dependent variable reflects collaborative output, not learning outcomes.

The largest effect for higher order thinking propensities is reported by Hu (2024), where students solved ethical dilemmas. The experimental group used a ChatGPT-assisted virtual learning companion, while one control group worked with a human peer and another worked individually. This comparison reflects plausible real-world scenarios, which makes for a meaningful comparison. Three outcomes were measured: self-rated problem-solving skills, self-reported motivational regulation, and self-rated ethical reasoning ability. While Deng et al. (2025) do not clarify which outcomes were designated higher-order thinking, only problem-solving and ethical reasoning fit the bill. Crucially, both outcomes rely on students' accurate self-reports, meaning higher-order thinking was stated but not demonstrated. Thus, these measures are unfit to assess actual higher order thinking skills due to learning.

Researchers aiming to conduct meaningful media comparison studies on tools like ChatGPT must impose constraints on their designs to ensure interpretable findings. To even get their project off the ground conceptually requires narrowing the research question from "What is the effect of ChatGPT on learning?" to, for example, "What is the effect of using ChatGPT in [X] way, with [Y] properties, compared to [Z] control condition?" Additional assumptions may still be needed, depending on how precisely the focal combination of ChatGPT's and instructional features is defined and operationalised. While meta-analyses can address broader questions, such as "What is the learning effectiveness of ChatGPT implementations in the literature?", the focus must still shift from ChatGPT as a tool to its role within specific instructional implementations. Meta-analysts must then carefully account for the diversity of treatments, controls, and outcomes.

5 | Auditing Primary Studies on Academic Performance

To document how often the emerging literature meets these three basic considerations, we audited the academic performance comparisons included in Deng et al. (2025). The audit draws on coding that is part of a larger systematic review of generative-AI studies in education (Lawson et al., forthcoming). With permission from the project partners, we report here the subset of codes needed to support the specific claims of this manuscript.

For each comparison, Lawson et al. (forthcoming) recorded (a) whether the ChatGPT treatment was described in sufficient detail to allow for a hypothetical replication, (b) whether the control-group activities were described at this same level of detail, and (c) whether the main outcome qualified as a learningmeasure–definedd as a non-self-report assessment administered after the intervention. Descriptions were rated *well-defined* when replication appeared feasible, *partial* when this was not present, but the description was more than merely labelling the group, and *not defined* when the study supplied little more than a label. Please note that, while the Forest plot (app B1 in Deng et al. 2025) contains 18 studies on academic performance, two compare three groups each, yielding 20 comparisons. However, as one of these compared nearly identical treatment groups, this

TABLE 1 | Audit summary.

Coding category	ChatGPT treatment	Control group	Outcome = learning
Well-defined	14/19 (74%)	8/19 (42%)	—
Partial	4/19 (21%)	8/19 (42%)	—
Not defined	1/19 (5%)	3/19 (16%)	—
Learning measure present	—	—	10/19 (53%)

did not warrant separate coding. A summary of results for 19 comparisons is reported (see Table 1).

Notably, only four comparisons (21%) satisfied all three criteria: Johnson et al. (2024), Meyer et al. (2024), Song and Song (2023), and Wu et al. (2024). These studies therefore yield interpretable effects related to learning. However, determining whether the observed effects can be attributed to ChatGPT alone requires further insight into whether key instructional features—aside from the chatbot—were held constant across treatment and control conditions. A more detailed analysis of this question will be presented in Lawson et al. (forthcoming).

6 | Thoughts on a Culture of Fast Science

Understandably, researchers want to get to the bottom of these questions: What is the potential role of AI in education? How significant are presumed benefits, and are they indeed transformational? Yet, alongside these intrinsic motivations, external incentives can drive the rush to publish hasty, preliminary work.[9] *Fast Science* (Frith 2020) generates research waste (Macleod et al. 2014), propagates errors (Smaldino and McElreath 2016), can lead to overload and anxiety (Bawden and Robinson 2009), and appears to deliver diminishing returns (Park et al. 2023). Robust science, on the other hand, begins with deeply thinking about the questions of interest, features clear definitions of concepts, emphasises attention to details, and is critical. Above all, robust science is cognizant of the importance of building a sturdy research base resistant to fleeting trends. While these ideals are universal (e.g., Merton 1973), the current rush to explore generative AI in education amplifies the urgency of slowing down. In the context of establishing learning effects of a novel technology, robust science means clearly defining the effect of interest, steering clear of well-documented conceptual pitfalls, and ensuring that methodological choices align with these goals rather than being dictated by convenience, trend-driven enthusiasm, or the limitations of an insufficiently developed literature.

What was true in the 1980s and 1990s is true today: "...we tend to ignore basic and applied research if that research was conducted with older media. We too often act as if we believe that each delivery technology requires a new theory of learning and performance. Thus, we 'reinvent the wheel' constantly but inadequately" (Clark 1994b, 8). As with technologies of the past, the allure of new technologies can lead to hasty research and unsound implementation, potentially undermining learning experiences for countless students globally. Educational technology researchers are tasked with identifying conditions of success and instructional pitfalls through research that holds up in the face of scrutiny. Without careful, deliberate research, we cannot hope to speak to policy, nor can we improve the lives of our students. Ultimately, robust science also means learning from the rich history of our field and the luminaries who shaped it to steer clear of seductive but entirely avoidable pitfalls.

Author Contributions

J. Weidlich: conceptualization, writing – original draft, writing – review and editing. **D. Gašević:** writing – review and editing, conceptualization. **H. Drachsler:** writing – review and editing, conceptualization. **P. Kirschner:** writing – review and editing, conceptualization.

Acknowledgements

We would like to thank Alyssa Lawson, Amedee Martella, Miriam Mulders, and Josef Buchner for their support in allowing us to use a slice of their data for this manuscript. Open access publishing facilitated by Universitat Zurich, as part of the Wiley - Universitat Zurich agreement via the Consortium Of Swiss Academic Libraries.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Endnotes

[1] We want to emphasise that our critique is grounded in a positivist paradigm, as it aligns with the methodological tradition of the studies we examine here. We acknowledge that other research traditions also offer valuable perspectives on learning with AI and ChatGPT, particularly in an emerging field where foundational understanding is still developing. Interpretivist and ethnographic approaches, for instance, explore how learners and educators make sense of AI in authentic settings, emphasising context and meaning (e.g., Biesta 2010; Jensen et al. 2022). Constructivist paradigms highlight the active role of learners in shaping their interactions with AI and may be better suited to studying these processes through constructivist rather than objectivist assumptions (Jonassen 1991). Finally, exploratory and design-based research approaches prioritise iterative, real-world intervention development over controlled hypothesis testing, making them particularly relevant for investigating how to effectively deploy new technologies for learning (Design-Based Research Collective 2003; Anderson and Shattuck 2012). While our focus remains on methodological rigour within experimental and quasi-experimental research, we recognise that diverse approaches contribute to a broader understanding of AI's role in education.

[2] Although the entire debate is fascinating, we particularly recommend reading the initial papers by Clark (1983), Kozma (1991), and

the summary by Tennyson (1994) to get an idea of the most essential points.

[3] Clark (1983) and Kozma (1991) are often cast as adversaries, yet their views are more complementary than contradictory. Kozma argues that a medium's affordances—its distinctive ways of representing information and supporting interaction—are inseparable from the instructional methods built upon them; together they shape how learners construct knowledge. This is an ontological claim about the nature of media-method relationships. Clark's argument conveys the methodological mirror image of that claim: when researchers attempt "medium-A versus medium-B" comparisons, they almost always alter the pedagogy as well, producing confounded designs and uninterpretable effects. This is an epistemological argument about what can be known from the available empirical evidence. And indeed, he showed that for cases where instructional activities are held constant and only the medium changes, no intrinsic learning benefit emerged. Thus, both positions converge in that meaningful effects arise from the medium-method package, not from the medium alone. If, however, the goal is causal inference about the technology only, it requires researchers to carefully "carve nature at its joints" by manipulating only one factor at a time (Shadish et al. 2002). That is, the instructional method must be specified and controlled. Then, any residual advantage of the treatment could be rightfully attributed to the unique properties of the technology.

[4] The only stringent interpretation from confounded media comparison studies with positive results is as follows: an opaque media-method combination was more effective for learning than (a) conventional teaching or (b) a similarly opaque sequence of events that explicitly did not include the media of interest. Note that neither interpretation allows us to single out that the media of interest was effective, thus failing the main research goal.

[5] Step-based tutoring provides hints and explanations on steps that students typically take when solving problems. Substep-based tutoring, a newer and more exacting approach, provides scaffolding and feedback at a finer level (Kulik and Fletcher 2016, 45)0.

[6] At this point, it should be noted that causal inference is, in principle, also possible from research designs without randomised assignment to treatment and control (i.e., quasi-experiments, natural experiments), and even for purely observational research. For the latter, however, causal inference requires specifying a (usually larger) set of assumptions under which the causal effect is identified. One increasingly popular way to model these assumptions is a graphical approach using Directed Acyclic Graphs (DAGs), pioneered by Pearl (2009). Primers for causal reasoning with DAGs for educational technology (Weidlich et al. 2024) and learning analytics (Weidlich et al. 2022) provide an introduction. For the sake of argument in this paper, we focus on causal inference without additional assumptions, which are usually more straightforward to defend via (quasi-)experiments.

[7] While the relative contribution of ChatGPT to the overall instructional approach could be derived from these codes, for a proper interpretation, there is still a need for a meaningful comparison.

[8] At this point, it is worth noting that Deng et al. (2025) do not refer to any of their outcome variables as 'learning'. Instead, they distinguish among variables like "academic performance" or "higher order thinking propensities". This distinction might appear to pre-empt concerns about the validity of the outcome as a measure of learning, given that performance outcomes could reflect students' effectiveness in using ChatGPT rather than genuine learning gains. Thus, the argument could go: Deng et al. (2025) did not aim to establish ChatGPT's effects on learning per se but rather a broader set of outcomes including both actual learning and performance measures plausibly related to learning. In our view, this terminological choice does not remove the need for scrutiny, as Deng et al.'s title, "Does ChatGPT enhance learning?", clearly frames the study around learning. This aim is also communicated in the conclusion, where the meta-analysis is said to address the knowledge gap surrounding the impact of ChatGPT on learning. Given this, we find it reasonable to discuss the meta-analysis and the associated literature on the grounds of wanting to establish effects on learning.

[9] Outlining the likely drivers of the prevalence of fast science is beyond the scope of this contribution, yet may include (a) incentive structures prioritising novel and exciting results over cumulative progress and truth (Nosek et al. 2012), (b) the increasing role of short-term third-party funding compared to long-term institutional funding (e.g., Wiener et al. 2020), (c) the undervaluing of peer review in academic publishing as a safeguard for research quality (e.g., Sculley et al. 2018), and others. Other factors, more unique to education and educational technology, include policy pressures, the need to produce immediately applicable research findings (Antoninis et al. 2023), and media attention and commercial influences proclaiming technology revolutions (e.g., Selwyn 2018).

References

Ahmed Moneus, A. M., and B. Q. Al-Wasy. 2024. "The Impact of Artificial Intelligence on the Quality of Saudi Translators' Performance." *Al-Andalus Journal for Humanities & Social Sciences* 11, no. 96: 202–230.

Anderson, J. R. 1982. "Acquisition of Cognitive Skill." *Psychological Review* 89, no. 4: 369–406.

Anderson, J. R., A. T. Corbett, K. R. Koedinger, and R. Pelletier. 1995. "Cognitive Tutors: Lessons Learned." *Journal of the Learning Sciences* 4, no. 2: 167–207.

Anderson, T., and J. Shattuck. 2012. "Design-Based Research: A Decade of Progress in Education Research?" *Educational Researcher* 41, no. 1: 16–25.

Antoninis, M., B. Alcott, S. Al Hadheri, et al. 2023. "Global Education Monitoring Report 2023: Technology in Education: A Tool on Whose Terms?" https://doi.org/10.54676/UZQV8501.

Baird, M. D., and J. F. Pane. 2019. "Translating Standardized Effects of Education Programs Into More Interpretable Metrics." *Educational Researcher* 48, no. 4: 217–228. https://doi.org/10.3102/0013189X19848729.

Bastani, H., O. Bastani, A. Sungu, H. Ge, O. Kabakcı, and R. Mariman. 2024. "Generative AI can harm learning" Available at SSRN,4895486. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4895486.

Bawden, D., and L. Robinson. 2009. "The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies." *Journal of Information Science* 35, no. 2: 180–191.

Bernard, R. M., P. C. Abrami, E. Borokhovski, et al. 2009. "A Meta-Analysis of Three Types of Interaction Treatments in Distance Education." *Review of Educational Research* 79, no. 3: 1243–1289.

Biesta, G. J. 2010. "Why 'What Works' Still Won't Work: From Evidence-Based Education to Value-Based Education." *Studies in Philosophy and Education* 29: 491–503.

Borenstein, M., L. V. Hedges, J. P. T. Higgins, and H. Rothstein. 2009. *Introduction to Meta-Analysis*. Wiley.

Buchner, J., and M. Kerres. 2023. "Media Comparison Studies Dominate Comparative Research on Augmented Reality in Education." *Computers & Education* 195: 104711.

Carpenter, S. K., A. E. Witherby, and S. K. Tauber. 2020. "On Students' (Mis)judgments of Learning and Teaching Effectiveness." *Journal of Applied Research in Memory and Cognition* 9, no. 2: 137–151. https://doi.org/10.1016/j.jarmac.2019.12.009.

Clark, R. E. 1983. "Reconsidering Research on Learning From Media." *Review of Educational Research* 53, no. 4: 445–459.

Clark, R. E. 1994a. "Media Will Never Influence Learning." *Educational Technology Research and Development* 42, no. 2: 21–29.

Clark, R. E. 1994b. "Media and Method." *Educational Technology Research and Development* 42: 7–10.

Clark, R. E. 2000. "Evaluating Distance Education: Strategies and Cautions." *International Journal of Educational Policy, Research, and Practice: Reconceptualizing Childhood Studies* 1, no. 1: 3–16.

Corbett, A. T., K. R. Koedinger, and J. R. Anderson. 1997. "Intelligent Tutoring Systems." In *Handbook of Human-Computer Interaction*, 849–874. North-Holland.

Darvishi, A., H. Khosravi, S. Sadiq, D. Gašević, and G. Siemens. 2024. "Impact of AI Assistance on Student Agency." *Computers & Education* 210: 104967.

Deng, R., M. Jiang, X. Yu, Y. Lu, and S. Liu. 2025. "Does ChatGPT Enhance Student Learning? A Systematic Review and Meta-Analysis of Experimental Studies." *Computers & Education* 227: 105224. https://doi.org/10.1016/j.compedu.2024.105224.

Design-Based Research Collective. 2003. "Design-Based Research: An Emerging Paradigm for Educational Inquiry." *Educational Researcher* 32: 5–8.

Deslauriers, L., L. S. McCarty, K. Miller, K. Callaghan, and G. Kestin. 2019. "Measuring Actual Learning Versus Feeling of Learning in Response to Being Actively Engaged in the Classroom." *Proceedings of the National Academy of Sciences* 116, no. 39: 19251–19257.

Eronen, M. I. 2020. "Causal Discovery and the Problem of Psychological Interventions." *New Ideas in Psychology* 59: 100785.

Fan, Y., L. Tang, H. Le, et al. 2024. "Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance." *British Journal of Educational Technology* 56: 489–530.

Frith, U. 2020. "Fast Lane to Slow Science." *Trends in Cognitive Sciences* 24, no. 1: 1–2.

Giannakos, M., R. Azevedo, P. Brusilovsky, et al. 2024. "The Promise and Challenges of Generative AI in Education." *Behaviour & Information Technology* 44, no. 11: 2518–2527.

Graesser, A. C., X. Hu, and R. Sottilare. 2018. "Intelligent Tutoring Systems." In *International Handbook of the Learning Sciences*, 246–255. Routledge.

Heckman, J., and R. Pinto. 2021. Econometric Causality: How to Express It and Why It Matters. https://cehd.uchicago.edu/wp-content/uploads/2021/04/Heckman_Pinto_2021_EconometricCausality.pdf.

Hernán, M. A., and J. M. Robins. 2020. *Causal Inference: What If*. Chapman & Hall/CRC.

Honebein, P. C., and C. M. Reigeluth. 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, no. 2: 465–496.

Hsu, M. H. 2024. "Mastering Medical Terminology With ChatGPT and Termbot." *Health Education Journal* 83, no. 4: 352–358.

Hu, Y. H. 2024. "Improving Ethical Dilemma Learning: Featuring Thinking Aloud Pair Problem Solving (TAPPS) and AI-Assisted Virtual Learning Companion." *Education and information technologies* 29, no. 17: 22969–22990.

Jensen, L. X., M. Bearman, D. Boud, and F. Konradsen. 2022. "Digital Ethnography in Higher Education Teaching and Learning—A Methodological Review." *Higher Education* 84, no. 5: 1143–1162.

Johnson, D. M., W. Doss, and C. M. Estepp. 2024. "Using ChatGPT With Novice Arduino Programmers: Effects on Performance, Interest, Self-Efficacy, and Programming Ability." *Journal of Research in Technical Careers* 8, no. 1: 1–17.

Jonassen, D. H. 1991. "Objectivism Versus Constructivism: Do We Need a New Philosophical Paradigm?" *Educational Technology Research and Development* 39: 5–14.

Jonassen, D. H., J. P. Campbell, and M. E. Davidson. 1994. "Learning With Media: Restructuring the Debate." *Educational Technology Research and Development* 42: 31–39.

Karpicke, J. D., and H. L. Roediger. 2008. "The Critical Importance of Retrieval for Learning." *Science* 319, no. 5865: 966–968.

Koedinger, K. R., and A. Corbett. 2006. "Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom." In *The Cambridge Handbook of The Learning Sciences*, edited by R. K. Sawyer, 61–77. Cambridge University Press.

Kozma, R. B. 1991. "Learning With Media." *Review of Educational Research* 61, no. 2: 179–211.

Kozma, R. B. 1994a. "A Reply: Media and Methods." *Educational Technology Research and Development* 42, no. 3: 11–14.

Kozma, R. B. 1994b. "Will Media Influence Learning? Reframing the Debate." *Educational Technology Research and Development* 42, no. 2: 7–19.

Kulik, J. A., and J. D. Fletcher. 2016. "Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review." *Review of Educational Research* 86, no. 1: 42–78.

Lawson, A. P., A. M. Martella, K. LaBonte, et al. 2024. "Confounded or Controlled? A Systematic Review of Media Comparison Studies Involving Immersive Virtual Reality for STEM Education." *Educational Psychology Review* 36, no. 3: 69.

Lawson, A. P., A. M. Martella, J. Weidlich, M. Mulders, and J. Buchner. forthcoming. "Color Me Confounded: A Critical Analysis of Media Comparisons on ChatGPT in Education."

Lehmann, M., P. B. Cornelius, and F. J. Sting. 2024. "AI Meets the Classroom: When Does ChatGPT Harm Learning?" *arXiv Preprint arXiv:2409.09047*.

Lou, Y., R. M. Bernard, and P. C. Abrami. 2006. "Media and Pedagogy in Undergraduate Distance Education: A Theory-Based Meta-Analysis of Empirical Literature." *Educational Technology Research and Development* 54, no. 2: 141–176.

Ma, W., O. O. Adesope, J. C. Nesbit, and Q. Liu. 2014. "Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis." *Journal of Educational Psychology* 106, no. 4: 901–918.

Macleod, M. R., S. Michie, I. Roberts, et al. 2014. "Biomedical Research: Increasing Value, Reducing Waste." *Lancet* 383, no. 9912: 101–104.

Mahapatra, S. 2024. "Impact of ChatGPT on ESL Students' Academic Writing Skills: A Mixed Methods Intervention Study." *Smart Learning Environments* 11, no. 1: 9.

Marshall, A. 1890. *Principles of Economics*. MacMillan and Company.

Merton, R. K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*, edited by N. W. Storer. University of Chicago Press.

Meyer, J., T. Jansen, R. Schiller, et al. 2024. "Using LLMs to Bring Evidence-Based Feedback Into the Classroom: AI-Generated Feedback Increases Secondary Students' Text Revision, Motivation, and Positive Emotions." *Computers and Education: Artificial Intelligence* 6: 100199.

Moher, D., L. Shamseer, M. Clarke, et al. 2015. "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement." *Systematic Reviews* 4: 1–9.

Morrison, G. R. 1994. "The Media Effects Question: 'Unresolvable' or Asking the Right Question." *Educational Technology Research and Development* 42: 41–44.

Niloy, A. C., S. Akter, N. Sultana, J. Sultana, and S. I. U. Rahman. 2024. "Is ChatGPT a Menace for Creative Writing Ability? An Experiment." *Journal of Computer Assisted Learning* 40, no. 2: 919–930.

Nosek, B. A., J. R. Spies, and M. Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability." *Perspectives on Psychological Science* 7, no. 6: 615–631.

Park, M., E. Leahey, and R. J. Funk. 2023. "Papers and Patents Are Becoming Less Disruptive Over Time." *Nature* 613, no. 7942: 138–144.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pigott, T. D., and J. R. Polanin. 2020. "Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review." *Review of Educational Research* 90, no. 1: 24–46.

Reiser, R. A. 1994. "Clark's Invitation to the Dance: An Instructional Designer's Response." *Educational Technology Research and Development* 42: 45–48.

Ross, S. M. 1994. "From Ingredients to Recipes... And Back: It's the Taste That Counts." *Educational Technology Research and Development* 42, no. 3: 5–6.

Salomon, G. 1990. "Cognitive Effects With and of Computer Technology." *Communication Research* 17, no. 1: 26–44.

Salomon, G. 1991. "Transcending the Qualitative-Quantitative Debate: The Analytic and Systemic Approaches to Educational Research." *Educational Researcher* 20, no. 6: 10–18.

Schunk, D. H. 2020. *Learning Theories: An Educational Perspective*. 7th ed. Pearson.

Sculley, D., J. Snoek, and A. Wiltschko. 2018. "Avoiding a Tragedy of the Commons in the Peer Review Process." *arXiv preprint arXiv:1901.06246*.

Selwyn, N. 2018. "Technology as a Focus of Education Policy." In *The Wiley Handbook of Educational Policy*, 457–477. Wiley.

Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton, Mifflin and Company.

Shrock, S. A. 1994. "The Media Influence Debate: Read the Fine Print, But Don't Lose Sight of the Big Picture." *Educational Technology Research and Development* 42: 49–53.

Smaldino, P. E., and R. McElreath. 2016. "The Natural Selection of Bad Science." *Royal Society Open Science* 3, no. 9: 160384.

Song, C., and Y. Song. 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.

Stadler, M., M. Bannert, and M. Sailer. 2024. "Cognitive Ease at a Cost: LLMs Reduce Mental Effort but Compromise Depth in Student Scientific Inquiry." *Computers in Human Behavior* 160: 108386.

Steenbergen-Hu, S., and H. Cooper. 2014. "A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on College Students' Academic Learning." *Journal of Educational Psychology* 106, no. 2: 331–347.

Sternberg, R. J., and K. Sternberg. 2017. *Cognitive Psychology*. 7th ed. Cengage Learning.

Tennyson, R. D. 1994. "The Big Wrench vs. Integrated Approaches: The Great Media Debate." *Educational Technology Research and Development* 42, no. 3: 15–28.

VanLehn, K. 2011. "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems." *Educational Psychologist* 46, no. 4: 197–221.

Wang, J., and W. Fan. 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, no. 1: 1–21.

Weidlich, J., D. Gašević, and H. Drachsler. 2022. "Causal Inference and Bias in Learning Analytics: A Primer on Pitfalls Using Directed Acyclic Graphs." *Journal of Learning Analytics* 9, no. 3: 183–199.

Weidlich, J., B. Hicks, and H. Drachsler. 2024. "Causal Reasoning with Causal Graphs in Educational Technology Research." *Educational technology research and development* 72, no. 5: 2499–2517.

Wiener, M., D. Maresch, and R. J. Breitenecker. 2020. "The Shift Towards Entrepreneurial Universities and the Relevance of Third-Party Funding of Business and Economics Units in Austria: A Research Note." *Review of Managerial Science* 14, no. 2: 345–363.

Wu, T. T., H. Y. Lee, P. H. Li, C. N. Huang, and Y. M. Huang. 2024. "Promoting Self-Regulation Progress and Knowledge Construction in Blended Learning via ChatGPT-Based Learning Aid." *Journal of Educational Computing Research* 61, no. 8: 3–31.

Yan, L., R. Martinez-Maldonado, Y. Jin, et al. 2025. "The Effects of Generative AI Agents and Scaffolding on Enhancing Students' Comprehension of Visual Learning Analytics." *Computers & Education* 234: 105322.

Zhang, L., P. A. Kirschner, W. W. Cobern, and J. Sweller. 2022. "There Is an Evidence Crisis in Science Educational Policy." *Educational Psychology Review* 34, no. 2: 1157–1176. https://doi.org/10.1007/s10648-021-09646-1.