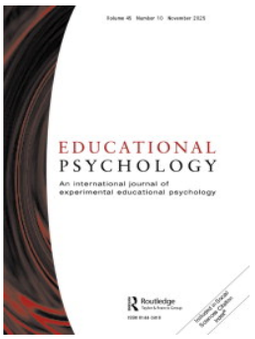

Routledge
Taylor & Francis Group

Educational Psychology

An International Journal of Experimental Educational Psychology

ISSN: 0144-3410 (Print) 1469-5820 (Online) Journal homepage: www.tandfonline.com/journals/cedp20

How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions

Rogers Kaliisa, Kamila Misiejuk, Sonsoles López-Pernas & Mohammed Saqr

To cite this article: Rogers Kaliisa, Kamila Misiejuk, Sonsoles López-Pernas & Mohammed Saqr (24 Sep 2025): How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions, Educational Psychology, DOI: 10.1080/01443410.2025.2553639

To link to this article: https://doi.org/10.1080/01443410.2025.2553639

Published online: 24 Sep 2025.

EDUCATIONAL PSYCHOLOGY
https://doi.org/10.1080/01443410.2025.2553639

Routledge
Taylor & Francis Group

How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions

Rogers Kaliisa[a], Kamila Misiejuk[b], Sonsoles López-Pernas[c] and Mohammed Saqr[c]

[a]Department of Education, University of Oslo, Oslo, Norway; [b]Centre for the Science of Learning & Technology (SLATE), University of Bergen, Bergen, Norway; [c]School of Computing, University of Eastern Finland, Joensuu, Finland

ABSTRACT

This exploratory meta-analysis synthesises current research on the effectiveness of Artificial Intelligence (AI)-generated feedback compared to traditional human-provided feedback. Drawing on 41 studies involving a total of 4813 students, the findings reveal no statistically significant differences in learning performance between students who received AI-generated feedback and those who received human-provided feedback. The pooled effect size was small and statistically insignificant (Hedge's $g = 0.25$, CI [−0.11; 0.60]), indicating that AI feedback is potentially as effective as human feedback. A separate meta-analysis focusing exclusively on studies in the domain of language and writing confirmed similar findings, with high heterogeneity persisting ($I^2 = 95\%$). The study further explored differences in feedback perception and found a small, negative, and statistically insignificant effect size (Hedge's $g = -0.20$, CI [−0.67; 0.27]). The study advocates for a hybrid approach, leveraging the scalability of AI while retaining the deep, empathetic, and contextual features of human feedback.

ARTICLE HISTORY
Received 15 May 2024
Accepted 25 August 2025

KEYWORDS
Artificial intelligence-feedback; human-feedback; learning analytics; meta-analysis

Introduction

Feedback as a tool to support learning has received significant attention in education. Hattie and Timperley (2007) define feedback as information an agent (e.g. teacher, peer, book, parent) provides regarding one's performance or understanding. This information aims to bridge the gap between what is understood and what is aimed to be understood, guiding students towards achieving specific learning goals. Studies have shown that delivering feedback appropriately and promptly can improve students' learning experiences and outcomes (Hattie & Timperley, 2007). However, with increasing enrolments in online and face-to-face learning environments, providing timely and appropriate feedback to large cohorts of students becomes

CONTACT Rogers Kaliisa rogers.kaliisa@iped.uio.no Department of Education, University of Oslo, Oslo, Norway

difficult, if not impossible, for teachers or peers. Where teachers and students use educational technologies, automated and artificial intelligence (AI)-assisted feedback systems powered by advanced techniques offer the potential to provide timely, personalised, and data-driven feedback to students, allowing for timely interventions and corrections. Such real-time responsiveness can enhance the learning experience, as students are provided with actionable assessments that can be immediately incorporated into their study strategies (González-Calatayud et al., 2021). For instance, the rapid advancement in AI-supported writing evaluation technologies, such as grammar checkers, style analysers, and AI tutors (Weitekamp et al., 2020), and large language models, such as ChatGPT (OpenAI, 2022), has influenced pedagogical practices by providing personalised, scalable, and immediate feedback (Bearman et al., 2023).

While AI and human feedback each have their own theoretical and empirical support, the comparative effectiveness of these feedback types on students' learning outcomes has not been comprehensively explored, especially in a meta-analytical format that considers varied educational contexts and diverse student populations. Moreover, as educational institutions increasingly integrate AI tools, assessing their effectiveness on student learning processes and outcomes becomes crucial. With this background, building on the existing studies and the foundational work of Hattie and Timperley (2007), this meta-analysis study explores how feedback provided by AI and humans impacts students' learning outcomes (e.g. performance, motivation, and satisfaction). Firstly, it will investigate the disparities in learning outcomes (e.g. performance) between students who receive AI-generated feedback and those who are guided by traditional teacher or peer feedback, as well as student perception of different types of feedback. Previous research has shown that peer and teacher feedback can differ in characteristics, student perception, or even effectiveness due to the differences in expertise between the two groups (e.g. Hamer et al., 2015; Pirttinen & Leinonen, 2022; Ruegg, 2015). As such, this comparison examines whether AI feedback aligns more closely with expert (teacher) feedback or peer feedback, or if it introduces entirely new dynamics in the perception and impact of feedback.

Secondly, the study will explore the psychological dimensions, comparing how the two modes of feedback impact students' motivation, engagement, and self-regulation. This study is particularly timely as AI-powered feedback systems become more prevalent, especially those based on large language models, such as ChatGPT, which warrants an understanding of how such technology can be harnessed effectively alongside traditional feedback systems (e.g. teacher and peer-supported) to improve student learning outcomes.

The rest of the paper is organised as follows. In the next section, we review relevant literature on AI feedback. This is followed by a presentation of the theoretical frameworks on learning and feedback, highlighting how they informed our understanding of AI and the effectiveness of human feedback. The section that follows presents existing meta-reviews on feedback and their gaps to justify the relevance of the current study. Later, the methodology, findings, and discussion of the findings are provided. We conclude the paper with the study's implications for teaching practice and future research.

Theoretical background

This section draws on three theoretical models to provide a framework for understanding the effectiveness of AI and human feedback in educational settings. These models are Hattie and Timperley's Feedback Model (2007), Kluger and DeNisi's Feedback Intervention Theory (1996), and the Self-Regulated Learning (SRL) framework by Butler and Winne (1995).

Hattie and Timperley's feedback model, formulated in their seminal 2007 paper, offers a detailed framework for understanding the mechanisms and impacts of feedback in educational settings. This model is particularly suited to the current meta-analysis, which compares the effectiveness of AI and human feedback on students' learning outcomes and dispositions. The model distinguishes feedback into three fundamental questions that should be addressed to optimise learning: 'Where am I going?' (highlighting its role in setting learning targets, Feed Up), 'How am I going?' (assessing current performance against these targets, Feedback), and 'Where to next?' (providing guidance on actions needed to achieve or enhance understanding, Feed Forward). The model emphasises the cyclical nature of feedback and is flexible enough to evaluate both AI- and human-provided feedback. It enables the assessment of how each type of feedback informs students about their current performance (e.g. writing, vocabulary) in relation to learning goals and what steps they might take next to improve. For example, the 'Focus on Feed Forward' aspect of the model is crucial for understanding the potential advantages of AI systems, which can potentially deliver more personalised, timely, and actionable feedback compared to traditional methods.

Kluger and DeNisi's Feedback Intervention Theory (FIT), developed in 1996, offers another critical lens through which to examine the effectiveness of feedback interventions. FIT posits that feedback impacts performance by directing learners' attention to different levels of behavioural regulation: task-specific processes, task motivation, and self-related processes. According to FIT, feedback is most effective when it focuses on task-specific processes and task motivation, as this directs attention towards actionable improvements and sustained effort. Conversely, feedback that shifts attention to self-related processes, such as self-esteem or general self-perceptions, often detracts from learning and can lead to reduced performance. This theory offers insights into how the delivery of feedback, whether by AI or humans, may influence the locus of learners' attention. For instance, AI systems might excel at maintaining task-focused feedback by avoiding emotionally charged responses that could trigger self-related processes. In contrast, human feedback might offer motivational benefits by fostering connection and engagement.

Complementing Kluger and DeNisi's and Hattie and Timperley's feedback models is the framework of Self-Regulated Learning (SRL), by Butler and Winne (1995). SRL emphasises the learner's active role in monitoring, evaluating, and adjusting their learning processes. Feedback plays a pivotal role in this framework by catalysing self-regulation. Through internal feedback mechanisms (self-monitoring) and external feedback sources (e.g. teachers or AI), learners assess their progress and refine their strategies. Effective feedback, according to the SRL model, should not only provide information on task performance but also support the development of self-regulatory skills, such as goal-setting, strategic planning, and self-evaluation. This perspective

emphasises the importance of feedback in fostering learner autonomy, a dimension where AI feedback might offer consistency and immediacy, while human feedback might excel in addressing individual motivational and contextual nuances.

Related literature: AI-generated feedback

The advancement of AI has led to the widespread use and adoption of automated and pre-trained large language models, such as the Generative Pre-Trained Transformer (GPT) (OpenAI, 2022) and the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). These models, trained on vast datasets, utilise a transformer architecture to generate human-like responses in response to prompts (Kasneci et al., 2023). One key development in education is the use of such models powered by advanced natural learning processing techniques to offer students timely, personalised, and data-driven feedback. Real-time responsiveness can enhance the learning experience by providing students with actionable insights that can be immediately incorporated into their study strategies (Kasneci et al., 2023). For instance, Escalante et al. (2023) investigated the use of generative AI tools, such as ChatGPT, for providing feedback to English as a New Language students and how they compare to human tutor feedback. The study found no significant differences in learning outcomes between students who received AI-generated feedback and those who received feedback from human tutors, suggesting that AI-generated feedback can be as effective as traditional methods without compromising educational quality. Guo and Wang (2024) explored the role of ChatGPT in assisting teachers with feedback on writing for English as a Foreign Language (EFL). The study involved a group of Chinese EFL teachers who compared feedback generated by ChatGPT with their own on students' argumentative essays. The findings indicated that ChatGPT provided significantly more feedback across content, organisation, and language aspects. The type of feedback also varied, with ChatGPT delivering more directive and praise-oriented comments, while teachers often provided more informative and questioning feedback. These findings suggest that while AI models like ChatGPT can enhance the quantity and promptness of feedback, they should be used in conjunction with human judgement to ensure the relevance and effectiveness of the feedback. Similar sentiments are shared in another recent work by Misiejuk et al. (2024), who employed GPT-3 to code student-generated content in online discussions based on intended learning outcomes. Findings revealed that while AI-supported coding is efficient, achieving substantial, moderate agreement with human coding for specific, nuanced, and context-dependent codes is challenging, suggesting a hybrid approach that integrates human judgement.

Besides the large language models, there are several AI-supported automated feedback systems, particularly for Automated Writing Evaluation (AWE), such as Criterion (Li et al., 2015), Pigai, Grammarly, AcaWriter (Knight et al., 2020) that provide formative feedback on different aspects of writing (e.g. grammar, structure). Li et al. (2015) used mixed methods to investigate how Criterion affected writing instruction and performance. Results suggested that Criterion led to increased revisions and that the corrective feedback provided by Criterion helped students improve accuracy from the initial to the final draft. Ding and Zou (2024) offer a comprehensive review of three prominent AWE tools (Grammarly, Pigai, and Criterion). The review suggests that

these AWE tools improve students' writing proficiency, particularly in grammar and syntax, although effectiveness may vary depending on the educational context and user engagement. However, even though AWE tools have been found useful in providing real-time feedback and personalised guidance, they are criticised for being less interpretable and biased, as the reasoning behind their decisions is often unclear to the users (Kasneci et al., 2023).

Several studies have explored student perceptions of AI-generated feedback in educational settings. These studies offer insights into how students perceive, interpret, and react to feedback provided by AI-assisted systems. For example, Ding and Zou (2024) reported that students view AWE systems positively, appreciating the immediate feedback and detailed error analysis provided by these systems. Escalante et al. (2023) investigated student preferences between AI-generated and human-generated feedback. The results indicated a split preference, with some students valuing the immediacy and precision of AI feedback, while others preferred the personal interaction and contextual insights offered by human feedback. Guo and Wang (2024) reported mixed reactions from teachers on the use of AI-assisted feedback, appreciating the efficiency and detailed nature of ChatGPT's feedback and the potential for reducing workload, yet concerned about the relevance and personalisation of the feedback, especially given the AI's lack of contextual understanding of the student's specific learning environment.

Previous meta-reviews

Previous meta-analyses have explored various aspects of feedback in education, providing insights into how different types of feedback affect learning outcomes. One of the early studies was by Azevedo and Bernard (1995), who analysed the impact of feedback on computer-based instruction. The study distinguishes between corrective feedback, which addresses errors or incorrect responses, and elaborative feedback, which provides explanations and extended information beyond mere correction. Effect size calculations from 22 studies involving the administration of immediate achievement posttests resulted in a weighted mean effect size of .80. Also, a mean weighted effect size of .35 was obtained from nine studies involving delayed posttest administration. The study revealed that elaborative feedback significantly improves learning outcomes due to its richer informational content. Zhai and Ma (2023) conducted a meta-analysis assessing the effectiveness of AWE on writing quality. They synthesised results from 26 studies with 2,468 participants between 2010 and 2022. The findings showed that AWE has a significantly positive effect on writing quality ($g = 0.861$). Moderator analyses reveal that AWE is more effective for post-secondary students than secondary students and shows greater benefits for English as a Foreign Language (EFL) and English as a Second Language (ESL) learners than native English speakers.

Ngo et al. (2024) focused on EFL/ESL learners, employing a three-level meta-analysis to investigate the impact of AWE on writing skills. Using a sample of 24 primary studies for between-group effects and 34 studies for within-group effects, the results revealed a medium overall between-group effect size and a large within-group effect size, indicating that AWE tools are generally effective in enhancing EFL/ESL writing performance. Fleckenstein et al. (2023) conducted a multi-level meta-analysis on the

effect of automated feedback on writing skills. Based on 20 studies involving 2,828 participants, the study reports a medium overall effect size ($g = 0.55$) of AWE on enhancing student writing performance. The analysis revealed significant heterogeneity among the included studies, indicating that AWE tools have a non-uniform effect on all learners or settings. Kluger and DeNisi (1996) offered a broader perspective, analysing feedback interventions involving 23,663 observations across various domains. Contrary to the notion that feedback enhances performance, this study found that over one-third of feedback interventions could decrease performance, underscoring the complex nature of feedback effects. This finding prompts further investigation into how automated and AI-assisted feedback, which is often highly task-focused, performs in comparison to more personalised human feedback.

One of the most comprehensive meta-analyses on feedback is by Wisniewski et al. (2019). This study analysed 435 studies with a collective sample size of over 61,000 participants. The study employed a random-effects model to address the significant heterogeneity observed in feedback effects based on timing, type, and delivery method. The results indicate a medium overall effect size ($d = 0.48$) of feedback on student learning, confirming feedback as a crucial component of effective teaching strategies. However, the study highlights that the effectiveness of feedback varies significantly and is influenced by factors, such as the content of the feedback and the learning domains it targets. For example, cognitive and motor skills benefitted more from feedback than motivational and behavioural outcomes.

Study motivation

The previous meta-analyses provide valuable insights into the role and impact of feedback across different educational settings and domains. However, existing studies still reveal a gap in the comparative analysis of AI *versus* human feedback across diverse learning outcomes. Recent developments in AI and the emergence of AI-assisted feedback based on sophisticated large language models have established a new reality that differs from past research, necessitating a robust synthesis of existing evidence. In that, a considerable number of our studies ($n = 29$, 70%) were published in the last 2 years, and around a third of the papers were published in the last year; therefore, they have not been fully covered by previous meta-analyses.

Furthermore, past synthesis research has focused either on specific aspects of learning (e.g. writing or pronunciation), learner populations (e.g. EFL/ESL students), or specific types of feedback (e.g. writing quality), without a broad comparison of feedback modalities across various educational levels and disciplines. For instance, Azevedo and Bernard (1995) emphasised the variability in feedback effectiveness depending on context and delivery methods. Such findings suggest that AI-supported feedback might offer advantages in consistency and scalability that human feedback cannot match, or vice versa. Moreover, the findings from Kluger and DeNisi (1996) regarding the potential negative impacts of poorly implemented feedback interventions underscore the need for a deeper understanding of how AI tools compare to human-provided feedback. This current study aims to fill these gaps by evaluating the effectiveness of AI feedback *versus* human feedback, considering various educational outcomes. This analysis is crucial for informing educational practices and policy,

particularly as AI technologies become increasingly prevalent in educational settings worldwide. The study intends to answer the following research questions:

RQ1: Do AI- and human-provided feedback affect students' learning performance differently?

RQ2: How does the perception of feedback differ between AI and human-provided feedback?

RQ3: Do AI and human-provided feedback affect students' motivation, engagement, and self-regulation differently?

RQ4: To what extent does hybrid feedback affect feedback perception, learning performance, and learning dispositions compared to exclusively AI or human-provided feedback?

Methodology

This study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework, updated by Hansen et al. (2022) and Page et al. (2021). In the following section, we describe the steps taken and the statistical approaches used to identify, code, and analyse the effect sizes from the included studies.

Literature search and inclusion criteria

On January 9, 2024, we searched three scientific databases relevant to our research questions: Scopus, Web of Science, and ERIC. We used the following search string in the title, abstract and author keywords: ('feedback') AND ('large language model' OR 'ChatGPT' OR 'natural language processing' OR 'artificial intelligence' OR 'automated' OR 'machine learning' OR 'rule-based' OR 'technology enhanced') AND ('education' OR 'students'). The search yielded 3,962 articles from Scopus, 785 from Web of Science, and 763 from ERIC (see Figure 1). In addition, to ensure we capture all previous studies, we included all articles listed in 25 previous meta-analyses that studied the effect of feedback on student learning ($n = 435$), as compiled by Wisniewski et al. (2019).

After removing duplicates, the remaining number of articles was 4,730. Three authors collaborated to export and sort these results using a web-based review software, Rayyan-ai (Johnson & Phillips, 2018). The search results were screened for relevance to the topic and quality based on the titles and abstracts. If the content of a paper was not reported in an abstract and a title, we referred to the full text. First, we excluded studies that were not situated in an educational setting. Moreover, we excluded studies that did not compare the effects of AI feedback with feedback given by a human assessor (peer or teacher). The included studies had to be (1) empirical studies, (2) contain quantitative data enough for the effect size synthesis, (3) study methodology is experimental or quasi-experimental, (4) be written in English, and (5) be published in peer-reviewed venues. After the first round of screening papers, 115 articles remained, and 4,615 were excluded.

The second round included the full-text reading of 115 included papers and their detailed coding. Papers included in this round were split among three researchers

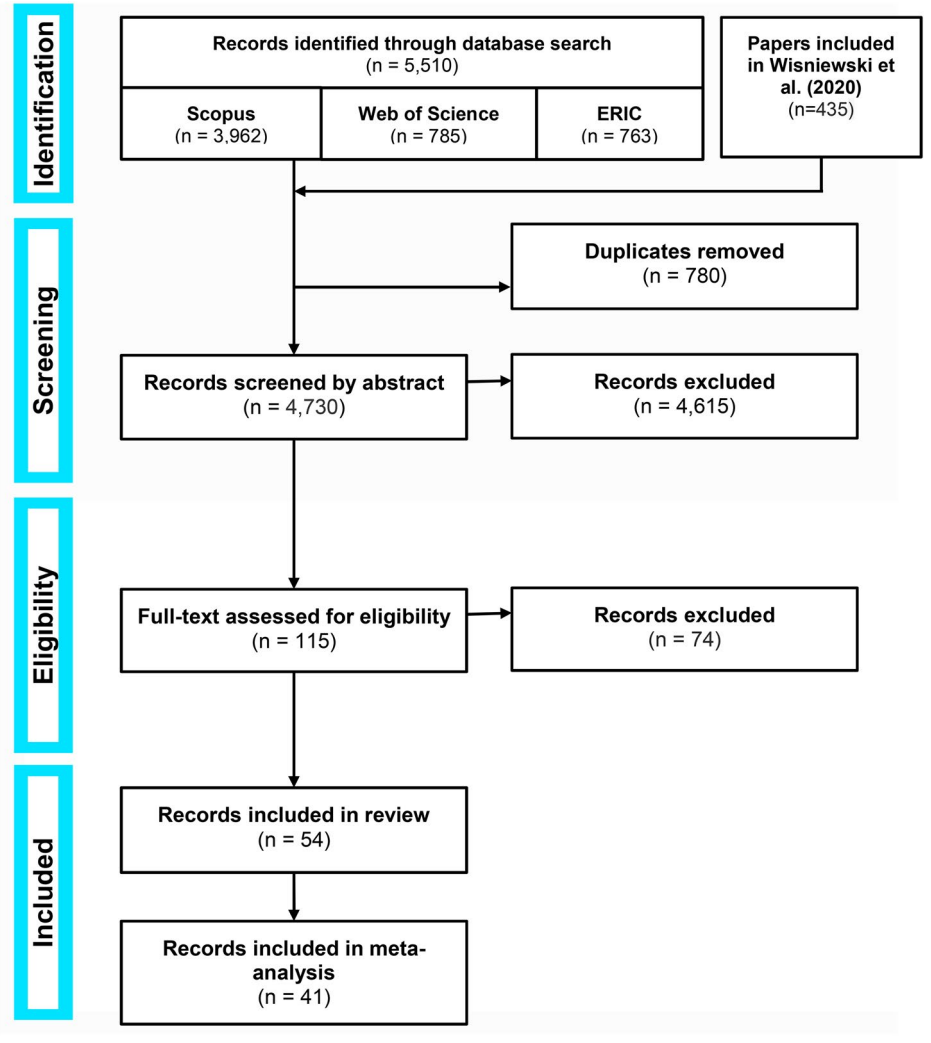

Figure 1. PRISMA flow diagram of search, inclusion, and exclusion screening.

for coding. In cases where the coding or inclusion of a paper was uncertain, multiple researchers read the full paper to extract relevant details. We retrieved information about the level of education, discipline, type of human feedback (peer or teacher), and type of AI feedback (only AI or hybrid). In addition, we coded the study design (experimental, quasi-experimental, pre-posttest, etc.) and outcome indicators: (1) *learning performance*, which focused on skills improvement or learning gains; (2) *feedback perception* indicating student perception of feedback received, and (3) *learning dispositions*, such as learning attitude, motivation, engagement, self-regulation, or self-efficacy. The second round of screening yielded 54 articles that met our inclusion criteria. After assessing the quality of each paper, we extracted relevant statistics, including sample sizes, standard deviations, means, and effect sizes (if available), for each outcome indicator. Two researchers double-checked the values extracted from the papers. Out of the 54 initially included studies, 13 articles were excluded since

they did not report all the necessary statistics for meta-analysis. Therefore, 41 papers were finally included.

Meta-analysis

Our coded dataset contained 41 articles that differed in their design and reported outcomes. Among those, 17 articles reported **multiple measures** per outcome (e.g. number of errors, vocabulary use, grammar, etc.). A total of 14 articles reported **multiple outcomes** (learning gains, feedback perception, engagement increase, etc.). Furthermore, 30 articles included **a repeated-measures design** (e.g. pre-post tests) comparing outcome measures between the control and experimental groups. Lastly, four articles reported **multiple independent samples** within the same article. We took several steps, as described below, to account for the nestedness and multiplicity of measures and study designs.

For studies that report **multiple measures** per outcome (e.g. for performance: grades for style, vocabulary, and grammar), researchers can either select a random measure of the reported outcomes, which risks biasing the results and reducing power, or pool all the reported outcomes using an appropriate method. We pooled all measures for a given sample using inverse-variance weighting, while accounting for the correlation between outcomes (i.e. using generalised least squares) (Pustejovsky & Tipton, 2022).

To pool the effect sizes, we calculated the effect size for each study (standardized mean difference, [SMD] between the experimental (AI feedback) and control (human feedback) groups of each measure per study. We used the *metafor* R package (Viechtbauer, 2010), providing the mean, standard deviation, and sample size of the control and experimental groups extracted from the studies for each measure. As a result of this step, we obtained a single effect size (SMD) and the corresponding sampling variance for each measure and study. We then reversed the direction of the effect size of those measures that were computed initially in a way that a higher value was associated with a worse outcome (e.g. the number of errors). All effect sizes were positive, indicating that the experimental group outperformed the control group. Following this step, we aggregated the effect sizes of all the measures that represented the same outcome using the methods mentioned above. This step resulted in each study outcome having a single effect size (SMD), either calculated from the overall score reported by the authors or pooled from the multiple scores.

Given the different nature of each outcome, we conducted separate meta-analyses for each outcome and study design to account for the multiple outcomes and the difference in study design (single-measure *vs.* repeated measures). As such, we conducted three meta-analyses: one on performance (distinguishing between task performance—single measure—and learning gains—repeated measures), and another on feedback perception (single measure). Feedback perception was operationalised as learners' self-reported ratings of the usefulness, clarity, relevance, or trust in the feedback, typically measured using Likert-scale questionnaires. The remaining outcomes and study designs did not include enough studies to conduct a meta-analysis. Moreover, for learning gains and feedback perception, we conducted separate meta-analyses for studies related to writing and language, as they constituted the

majority of the studies that allowed for individualised inspection. We conducted multilevel meta-analyses to account for the nested independent studies per article in each meta-analysis. We specified that studies were clustered per article using the *rma.mv* function from the *metafor* R package. The input for each meta-analysis was the effect size (SMD) and sampling variance of the outcome for each study as provided by the *aggregate.escalc* function. The SMD was computed using Hedge's *g*, whereby an effect size of 0.2 is considered small, 0.5 medium, and 0.8 large.

We used a random-effects model to pool the effect size estimates of the included studies, as we expected a high level of heterogeneity, which was later confirmed by the I^2 statistic (Higgins & Thompson, 2002). Given that our meta-analysis had a complex structure, possible dependence of outcomes, and clusters of samples within the same study, we used Robust Variance Estimation—with sandwich-type estimator—to account for effect size dependence and avoid underestimation of variance, inflation of confidence intervals or Type I error that may arise from using traditional models (Hedges et al., 2010; Pustejovsky & Tipton, 2022). The Cochrane Handbook for Systematic Reviews of Interventions (Green & Higgins, 2011) states that an I^2 lower than 0.4 represents negligible heterogeneity, an I^2 between 0.3 and 0.6 is moderate, substantial between 0.5 and 0.9, and considerable between 0.75 and 1. We estimated the between-study variance using the restricted maximum-likelihood estimator (RMLE) (Viechtbauer, 2010), as recommended for continuous outcomes by recent guidelines (Veroniki et al., 2016). We investigated potential moderators and sources of heterogeneity among the included studies by examining potential moderators, including year of publication as an indicator of AI technology level and type of feedback (teacher or peer).

Meta-analyses may be influenced by outliers or publication bias. We present a comprehensive analysis of outliers, influence, and bias sources in the Appendix (see Tables A2–A4), and provide the results here concisely. To investigate outliers, we searched for studies where either the lower or upper limit of the confidence interval lay outside the confidence interval of all studies. We identified extremely small effect sizes—where the effect size lies below the lower bound of the CI of the pooled effect size—and extremely large effect sizes when the effect size is larger than the upper bound of the CI of the pooled effect size. For each meta-analysis, we examined the presence of outliers, re-analysed the results after removing outliers, and reported the changes (Viechtbauer & Cheung, 2010).

To further assess the influence of individual studies on the overall meta-analytic results, we conducted a leave-one-out sensitivity analysis. This approach involves iteratively excluding each study from the meta-analysis and recalculating the pooled effect size to observe how the exclusion impacts the results. Significant changes in the pooled effect size or confidence intervals after removing a study indicate that the study has a substantial influence on the meta-analysis. Identifying such influential studies enables the identification of potential sources of heterogeneity and assess the robustness of the meta-analytic findings (Viechtbauer & Cheung, 2010). Furthermore, we generated Graphic Display of Heterogeneity (GOSH) plots. GOSH plots involve resampling subsets of studies and plotting the resulting pooled effect sizes to visualise the distribution and patterns of heterogeneity within the data. These plots help identify clusters of studies with similar effect sizes, outliers, and potential sources of variability (Olkin et al., 2012).

To rigorously investigate publication bias, we investigate bias both visually and statistically. Visually, a funnel plot of observed effect sizes (on the x-axis) and their standard error (on the y-axis) was plotted. In the absence of publication bias, the data points of the effect sizes should be symmetrical. Statistically, such asymmetry can be tested using Egger's regression analysis test, which quantitatively estimates the symmetry of the funnel plot. Further, a Rank Correlation Test for Funnel Plot Asymmetry was also performed to test whether the observed effect sizes are correlated with the corresponding sampling variances. The high correlation indicates the funnel plot asymmetry and possible publication bias (Begg & Mazumdar, 1994). Moreover, we employed the trim-and-fill method to investigate further and adjust for publication bias. This method enhances the interpretation of funnel plot asymmetry by estimating the number of potentially missing studies that would make the plot symmetrical. It then imputes these missing studies and recalculates the pooled effect size, providing a more conservative estimate that accounts for the potential impact of publication bias on the meta-analytic findings (Duval & Tweedie, 2000).

Results

Descriptives of studies

Table A1 (see Appendix) shows the 41 included articles. In total, the studies have a combined sample of 4,813 students. Figure 2 shows the number of articles per educational field and level. The field of 'Language and writing' was the most common in our dataset, with 33 articles (29 in higher education, three in K-12, and one in open education). A total of six articles were conducted in the field of STEAM (Science, technology, engineering, arts, and mathematics) (one in K-12 and five in higher education). There was one article in sports education and one in social sciences. Learning performance (understood as knowledge acquisition or task performance) was the most commonly assessed outcome (32 articles), followed by feedback perception (12 articles). Other articles (12) reported on a variety of outcomes related to learning dispositions, including motivation, self-regulation, engagement, and emotions.

Additionally, we categorised the AI systems and types of feedback used across the included studies. Among the 41 articles, AI tools were grouped based on whether they employed rule-based systems ($n = 20$), machine learning-based tools ($n = 6$), or generative AI models ($n = 5$). Rule-based systems, such as Pigai (e.g. Shang, 2022;

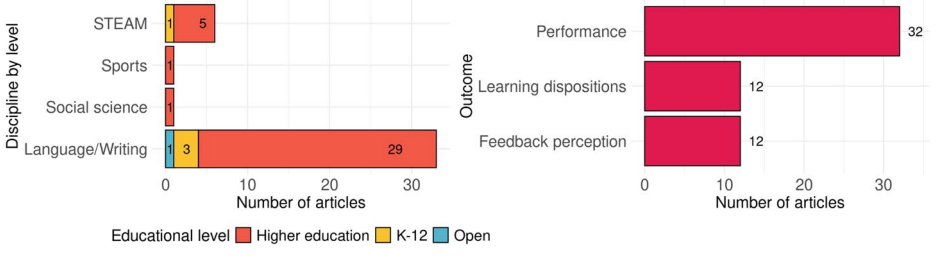

Figure 2. Left: Number of articles per discipline and educational level. Right: Number of articles that cover each outcome.

Wang & Han, 2022), Grammarly (e.g. Chang et al., 2021), and MY Access (e.g. Lai, 2010), typically provide directive, corrective feedback, often at the sentence or grammatical level. Some systems, such as Criterion (Hassanzadeh & Fotoohnejad, 2021) and iWrite (Chen & Pan, 2022), provide more comprehensive feedback on writing traits, including coherence and organisation. A smaller subset utilised more recent generative AI models, such as ChatGPT (e.g. Escalante et al., 2023; Silitonga et al., 2023), which provided conversational and content-reflective feedback. Other systems, such as BERT-based feedback platforms (Darvishi et al., 2022, 2024), utilise machine learning for feedback ranking and improvement suggestions. However, some studies did not specify the AI tool used or provide sufficient detail to classify them (e.g. Ouyang et al., 2023; Rosen, 2015; Ruwe & Mayweg-Paus, 2023; Xu et al., 2021).

Do AI- and human-provided feedback affect students' learning performance differently?

Two groups of studies assessed performance: the first group assessed the performance of AI feedback compared to human (peer or teacher) feedback (single-measure studies, 11 articles), and the second group assessed the change in performance after intervention with feedback (pre-post studies, 14 articles).

Task performance

The pooled effect size of the SMD in task performance (post-test only) between AI and human feedback in the 11 articles (12 studies) was small (Hedge's $g = 0.25$) and statistically insignificant (CI [−0.11; 0.60]). The heterogeneity was substantial ($I^2 = 75.0\%$ [56.0%; 85.8%]), and the prediction interval ranged from −0.87 to 1.37, indicating wide heterogeneity and uncertainty regarding effect sizes of future replications (see the forest plot in Figure 3). Across the included studies, there was significant variability in both the AI technologies and the types of feedback provided. For example, Studies using rule-based AWE systems like Pigai (Sun & Fan, 2022), MY Access (Lai, 2010), and Criterion (Hassanzadeh & Fotoohnejad, 2021) typically provided corrective feedback

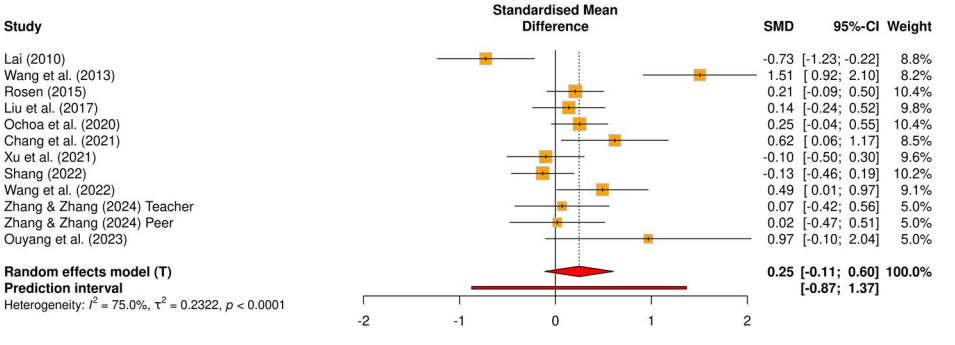

Figure 3. Forest plot of the meta-analysis of performance (post-assessment). A positive SMD indicates that AI feedback yielded higher learning performance, while a negative SMD suggests the opposite.

focused on grammar and syntax, while others, such as iWrite (Chen & Pan, 2022) and AcaWriter (Shibani et al., 2017), offered higher-level feedback related to writing structure and rhetorical quality. Some generative AI systems, such as ChatGPT (Escalante et al., 2023) and WordTune (Rad et al., 2023), were also utilised, providing directive, elaborative, and conversational feedback. Such differences in feedback types and types of AI models may account for the variability in performance outcomes observed across studies.

An outlier assessment identified two studies, Wang et al. (2013) and Lai (2010), as outliers. The pooled effect size of the SMD without outliers was slightly lower (Hedge's $g = 0.19$) but statistically significant (CI [0.01; 0.36]). Yet, the prediction interval was −0.11 to 0.48, indicating a wide range of uncertainty of future replications. Furthermore, Kendall's Rank correlation test, as described by Begg and Mazumdar (1994), for funnel plot asymmetry was statistically insignificant (Kendall's $\tau = 0.18$, $p = 0.46$), indicating no evidence of publication bias. Further, Egger's regression analysis for Funnel Plot Asymmetry was also statistically insignificant ($z = 1.02$, $p = 0.31$), corroborating the previous conclusions of an absence of evidence of publication bias. An in-depth analysis of sensitivity, bias, and influence using the trim-and-fill, leave-one-out, and GOSH methods (Table A1) suggests that outlier studies and publication bias may have a notable impact on the pooled effect size and heterogeneity in the performance meta-analysis. While the adjusted effect sizes (Hedge's $g = 0.09$–0.19) are smaller than the original estimate (Hedge's $g = 0.25$), they still indicate a small positive effect. However, the high levels of heterogeneity observed in several models suggest that further investigation into sources of variability across studies is warranted.

For this purpose, a subgroup analysis showed that the SMD in performance between AI and peer feedback was 0.00 (CI [−0.59; 0.58]) and 0.41 (CI [−0.10; 0.92]) for teachers. However, the test of moderators was statistically insignificant QM ($df = 1$) = 1.76, p = 0.18, and the test for residual heterogeneity was statistically significant and showed that the type of human feedback accounted for only 7.82% of the heterogeneity. Given that technological advances could affect how students perceive feedback delivered by AI and, consequently, the effect it may have, we tested the effect of the year as a moderator. The effect of the year was small, positive, but statistically insignificant according to the test of moderators: QM ($df = 1$) = 0.06, $p = 0.80$. The test for residual heterogeneity was statistically significant: QE ($df = 10$) = 44.03, $p < .0001$. Moreover, the study year accounted for 0% of the heterogeneity. Similarly, statistically non-significant results were obtained when testing the moderating effect of the study discipline: QM ($df = 1$) = 0.32, p-val = 0.57.

In summary, there is little or no evidence—considering outlier removal—that supports the conclusion that AI feedback has a better effect on students' performance than human feedback. The high heterogeneity and wide range of prediction intervals (from negative to positive values in both cases) make it unlikely that AI will yield any significant meaningful difference in future replications.

Performance gains

The second group of studies assessed the performance gain compared to the baseline (pre-post test) as a response to feedback by AI or humans. The pooled effect size of

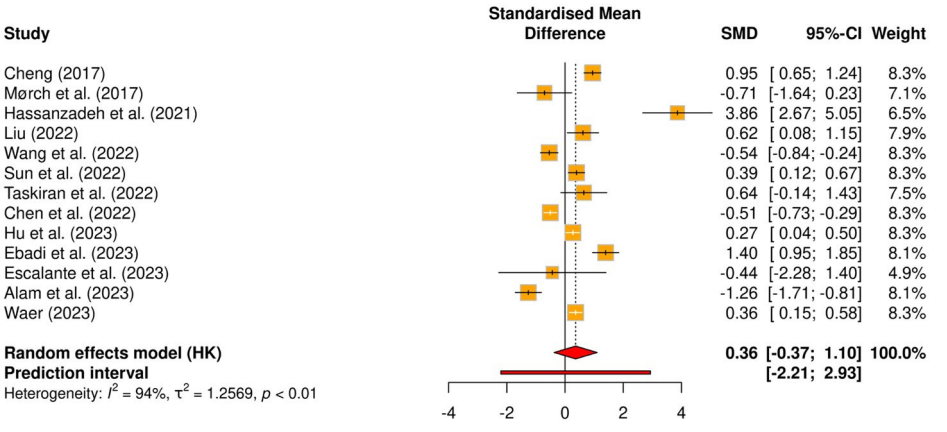

Figure 4. Forest plot of the meta-analysis of performance increase (post-assessment–pre-assessment). A positive SMD indicates that AI feedback yielded higher performance gains, and a negative SMD indicates otherwise.

the difference in performance increase was small (Hedge's $g = 0.36$) and statistically insignificant (CI [−0.37; 1.1]) between AI and human feedback (see Figure 4). The heterogeneity was considerable ($I^2 = 94\%$ [91.6%; 95.9%]), and the prediction interval ranged from −2.21 to 2.93, indicating very high heterogeneity and a wide variability in future replications. Specific AI systems, such as Virtual Writing Tutor (Mohammadi et al., 2023) and Pigai (Lai, 2010; Sun & Fan, 2022), have demonstrated advanced capabilities in offering rewriting suggestions and vocabulary expansion, which may contribute to the marginally improved performance observed in some studies. Meanwhile, some automated AI tools, such as 'Write & Improve' used by Taskiran and Goksel (2022), provided error identification without direct corrections, while generative tools like WordTune (Rad et al., 2023) offered detailed improvement suggestions or rewriting alternatives that preserved the original meaning. Such nuanced approaches, particularly those integrating direct suggestions, may help explain why some studies observed marginally better performance outcomes.

An outlier assessment identified two studies, Hassanzadeh and Fotoohnejad (2021) and Alam and Usama (2023), as outliers. Yet, no considerable change in effect size was observed (Hedge's $g = 0.28$) and was statistically insignificant (CI [−0.18; 0.73]). Similarly, the prediction interval was wide, −1.20 to 1.75, and the heterogeneity was considerable. Both publication bias tests (rank correlation and Egger's regression) were statistically insignificant, indicating no evidence of publication bias (Kendall's $\tau = 0.05$, $p = 0.86$; Egger's regression $z = 0.50$, $p = 0.62$). The results of the in-depth sensitivity, bias, and influence analysis (Table A2) suggest that the learning gains meta-analysis is characterised by high heterogeneity, with I^2 consistently exceeding 90% across all analyses (ranging from 92.1 to 94.0%). This high level of heterogeneity persists even after outlier removal or bias correction. The pooled effect sizes vary between analyses, from Hedge's $g = 0.13$ (after GOSH outlier removal) to Hedge's $g = 0.45$ (after leave-one-out influential cases removal), with wide confidence intervals in all cases, indicating considerable uncertainty in the effect size estimates. While outliers and influential studies contribute to

variability, they do not fully account for all the heterogeneity, begging the need for further exploration of study-level moderators or methodological differences that may explain the observed variability. A subgroup analysis showed that the SMD in performance gains between AI and peer feedback was −0.82 (CI [−1.86; 0.22]), and 0.72 (CI [−0.09; 1.54]) for teachers. The test of moderators was statistically significant QM ($df = 1$) = 4.82, $p = 0.03$, indicating an advantage of teachers' feedback over that of peers. The test for residual heterogeneity was statistically significant QE ($df = 11$) = 114.57, $p < .0001$, where the year accounted for 24.98% of the heterogeneity. Similarly, the year had no statistically significant effect on the magnitude of the difference according to the test of moderators: QM ($df = 1$) = 0.06, p-val = 0.80, and accounted for 0% of the heterogeneity. Lastly, statistically non-significant results were also obtained when testing the moderating effect of the study discipline: QM ($df = 1$) = 0.01, p-value = 0.93. However, several studies, such as Xu et al. (2021) and Hu et al. (2023), did not specify the underlying AI technologies used, making it difficult to fully assess how the sophistication or category of AI (e.g. rule-based *vs.* generative) may have influenced the feedback and subsequent learning outcomes.

In summary, there was no statistically significant difference in performance gain between AI and human feedback. Although a statistically significant difference was observed between teacher- and peer-provided feedback, favouring teachers, the effect size of each subgroup was still not statistically significant.

As all but one study that measured learning gains were in the field of language and writing, we conducted a separate meta-analysis after excluding Hu et al. (2023). The pooled effect size of the difference in performance increase was very similar to the one obtained for the complete pool (Hedge's $g = 0.38$) and statistically insignificant (CI [−0.44; 1.19]) between AI and human feedback with high heterogeneity ($I^2 = 95\%$), and the prediction interval ranged from −2.39 to 3.14, corroborating again the very high heterogeneity of uncertain and wide variability of future replications.

How does feedback perception differ between AI and human-provided feedback?

Effective feedback should not only be passively received but also acted upon (Iraj et al., 2021). In this study, we conceptualised feedback perception as students' subjective evaluation of the feedback they received, including dimensions, such as clarity, usefulness, and agreement with the feedback (e.g. Van der Pol et al., 2008; Wu & Schunn, 2020). This interpretation aligns with how feedback perception was operationalised in the included studies—for instance, through student surveys (Wilson & Martin, 2015), or interaction-based metrics, such as the number of likes or rate of likes given to feedback comments (Darvishi et al., 2022, 2024). The reviewed studies implemented a range of feedback types that may have influenced students' perception. For instance, directive feedback common in tools like Grammarly, Pigai, and iWrite typically provides prescriptive suggestions focused on grammar, structure, or lexical choice. In contrast, conversational feedback, enabled by tools like ChatGPT (e.g. Escalante et al., 2023; Silitonga et al., 2023), allowed students to engage in iterative dialogues around their work. Additionally, feedback mechanisms varied in terms of

immediacy and interactivity, with some systems offering real-time corrections (e.g. Criterion, Virtual Writing Tutor) and others providing summative evaluations after the task.

The pooled effect size of students' assessment of feedback perception of AI *versus* human was small, negative, and statistically insignificant (Hedge's $g = -0.20$, CI [−0.67; 0.27]). There was also considerable heterogeneity $I^2 = 84.7\%$ [72.7%; 91.4%], and the prediction interval ranged from −1.56 to 1.16. The conceptualisation of feedback perception varied significantly across studies and types of AI models. For instance, Generative AI models, such as ChatGPT (Escalante et al., 2023; Silitonga et al., 2023) enabled conversational feedback, while rule-based AI systems like Grammarly (Ebadi et al., 2023) provided corrective feedback at a granular level, focusing primarily on grammar. Similarly, tools like BERT-based systems (Darvishi et al., 2024) integrated additional features, such as identifying low-quality feedback and offering detailed improvement suggestions. Such variations highlight the evolving capabilities of AI tools and their influence on feedback perception.

Similar to the two previous meta-analyses, we conducted an in-depth bias, sensitivity, and influence analysis. The results (Table A3) show that, while the removal of outliers reduces heterogeneity, it remains relatively high ($I^2 = 75.4$–84.7%), suggesting that study-level differences or methodological inconsistencies contribute to variability. The results are, however, robust against publication bias, as indicated by the trim-and-fill method. Further analyses are needed to explore potential moderators or subgroup differences and refine the conclusions, thereby providing a better understanding of the high heterogeneity. For instance, the effect of year on the perception of feedback was statistically significant, with a positive estimate of 0.09 ($p = 0.03$), and the test of moderators was also statistically significant, QM ($df = 1$) = 4.73, $p = 0.03$. The year also accounted for 34.4% of the heterogeneity. Lastly, statistically non-significant results were obtained when testing the moderating effect of the study discipline: QM ($df = 1$) = 0.08, p-val = 0.78, which also accounted for 0% of the heterogeneity (Figure 5).

A large proportion of the studies analysing the difference in feedback perception depending on whether the feedback comes from humans or AI (all but two) focused on the field of language and learning. Therefore, we conducted a separate meta-analysis for this specific subset of studies. The results closely aligned with those obtained

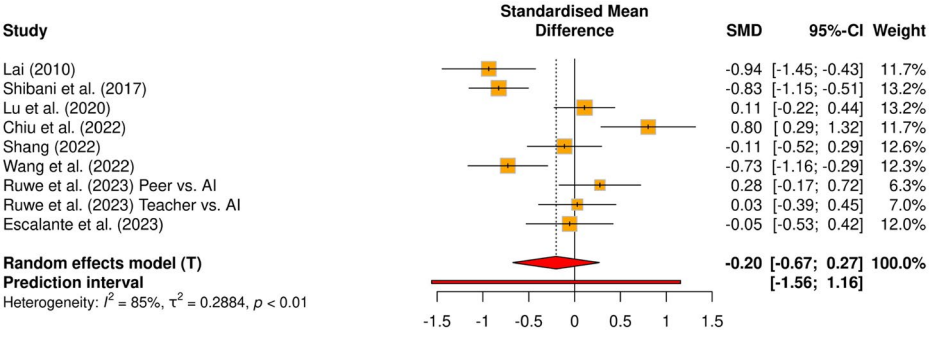

Figure 5. Forest plot of the meta-analysis of feedback perception. A positive SMD indicates that AI feedback was perceived more positively, and a negative SMD indicates otherwise.

from the complete data pool. The pooled effect size remained small and non-significant (Hedge's $g = -0.24$, 95% CI [−0.68, 0.21]), although heterogeneity was reduced ($I^2 = 74\%$). The confidence interval, however, remained wide (−1.35 to 0.87).

Do AI- and human-provided feedback affect students' motivation, engagement, and self-regulation differently?

Studies in this category (the effect of AI on dispositions) were rather diverse, measuring several outcomes with different measures, and therefore, were difficult to combine in a meta-analysis. Thus, they will be synthesised qualitatively. The majority of studies followed the direction of the previous meta-analyses and reported statistically insignificant effect sizes. In the remaining studies, the reported results were often contradictory, with some authors reporting a positive effect and others reporting the opposite. For instance, Chiu et al. (2022) reported a positive effect of AI feedback on learning attitude, while Lai (2010) reported a comparable negative effect size. We see the same contradiction even within the same article for peer and teacher feedback in terms of self-evaluation (Zhang & Zhang, 2024). Several systems conceptualised feedback as a tool for fostering deeper learning and engagement. For example, Chiu et al. (2022) integrated interventions that encouraged low-performing students to practice more intensely while providing supplementary materials for students who performed at a medium level. Similarly, Huang (2020) utilised Pigai to facilitate vocabulary expansion by suggesting synonyms for words in student writing, while Wilson and Martin (2015) employed PEG Writing, an AI-powered tool, to provide interactive multimedia materials for developing specific writing skills. Such designs highlight how AI feedback can contribute to broader learning dispositions beyond immediate task performance. In summary, results vary widely, with no strong evidence of positive effects, except for sporadic papers that have yet to be confirmed in future studies (Table 1).

To what extent does hybrid feedback affect feedback perception, learning performance, or learning dispositions compared to exclusively AI- or human-provided feedback

Only sixteen studies in our review reported the effect of human-only or AI-only feedback compared with a hybrid condition, where feedback was created through a synergy between a human and AI (see Table 2). Six out of eleven studies comparing the impact of teacher feedback with hybrid feedback provided by a teacher and AI reported a statistically significant positive effect of hybrid feedback on a post-test performance ($n = 1$), performance gains ($n = 3$), learning dispositions ($n = 1$), and engagement ($n = 1$). One study reported a statistically significant negative effect of a hybrid condition on performance gains. Three studies in our review compared peer-only feedback with a hybrid (Peer + AI) condition. Only one study reported statistically significant results, indicating a positive effect of a hybrid condition on feedback perception, specifically the number of likes a feedback comment received. Finally, two studies in our review examined the effect of the hybrid condition (Teacher + AI) against AI-only feedback. The results were statistically significant in one of the studies. They indicated a negative effect of AI-only feedback on the correct use of articles in English as a foreign language assignment.

Table 1. The pooled effect size of each learning disposition per study.

Study	Variable	Effect size
Chiu et al. (2022)	Learning attitude	0.66 (CI [0.06; 1.25])
Lai (2010)	Learning attitude	−0.71 (CI [−1.23; −0.20])
Ouyang et al. (2023)	Social engagement	0.36 (CI [−0.47; 1.20])
	Cognitive engagement	0.29 (CI [−0.55; 1.14])
Rosen (2015)	Behavioural engagement	0.02 (CI [−0.29; 0.33])
	Monitoring	0.54 (CI [0.19; 0.88])
	Motivation	0.00 (CI [−0.34; 0.34])
Ruwe and Mayweg-Paus (2023) Peer *vs.* AI	Motivation	0.36 (CI [−0.13; 0.86])
Ruwe and Mayweg-Paus (2023) Teacher *vs.* AI	Motivation	−0.02 (CI [−0.49; 0.45])
Ruwe and Mayweg-Paus (2023) Peer *vs.* AI	Self-efficacy	0.41 (CI [−0.09; 0.90])
Ruwe and Mayweg-Paus (2023) Teacher *vs.* AI	Self-efficacy	0.46 (CI [−0.01; 0.93])
Ruwe and Mayweg-Paus (2023) Peer *vs.* AI	Enjoyment	0.55 (CI [0.06; 1.05])
Ruwe and Mayweg-Paus (2023) Teacher *vs.* AI	Enjoyment	0.06 (CI [−0.41; 0.52])
Ruwe and Mayweg-Paus (2023) Peer *vs.* AI	Anger (reversed)	0.11 (CI [−0.38; 0.61])
Ruwe and Mayweg-Paus (2023) Teacher *vs.* AI	Anger (reversed)	−0.28 (CI [−0.75; 0.19])
Silitonga et al. (2023)	Motivation	0.67 (CI [0.62; 0.72])
Sun and Fan (2022)	Anxiety (reversed)	0.11 (CI [0.09; 0.14])
Waer (2023)	Apprehension	0.43 (CI [0.41; 0.45])
Zhang & Zhang (2024) Teacher	Learning strategies	−0.29 (CI [−0.31; −0.27])
Zhang & Zhang (2024) Peer	Learning strategies	−0.04 (CI [−0.06; −0.02])
Zhang & Zhang (2024) Teacher	Self-evaluation	−0.04 (CI [−0.08; −0.01])
Zhang & Zhang (2024) Peer	Self-evaluation	0.10 (CI [0.07; 0.13])

Table 2. Studies comparing hybrid and human-only or AI-only conditions.

Outcome	Study	Control	Experimental	Variable	Effect size
Performance (post-only)	Wilson and Czik (2016)	Teacher	Teacher + AI	Overall score	−0.05 (CI [−0.38; 0.27])
	Mohsen and Alshahrani (2019)	Teacher	Teacher + AI	Writing score	−0.84 (CI [−2.02; 0.34])
	Fan (2022)	Teacher	Teacher + AI	Syntactic complexity measure	1.07 (CI [0.70; 1.43])
Performance gain	Cheng (2022) I	Teacher	Teacher + AI	Essay scores	0.50 (CI [0.35; 0.65])
	Cheng (2022) II	Teacher	Teacher + AI	Essay scores	−0.21 (CI [−0.33; −0.09])
	Ebadi et al. (2023)	Teacher + AI	AI	Correct use of articles	−0.56 (CI [−0.80; −0.31])
	Ebadi et al. (2023)	Teacher	Teacher + AI	Correct use of articles	1.96 (CI [1.79; 2.12])
	Mohammadi et al. (2023)	Teacher + AI	AI	Writing score gain	−0.55 (CI [−1.14; 0.03])
	Mohammadi et al. (2023)	Teacher	Teacher + AI	Writing score gain	−0.59 (CI [−1.46; 0.28])
	Wang (2019)	Teacher	Teacher + AI	Writing score	0.10 (CI [0.07; 0.13])
Feedback perception	Darvishi et al. (2022)	Peer	Peer + AI	Number of likes	1.38 (CI [1.16; 1.61])
	Darvishi et al. (2024)	Peer	Peer + AI	Rate of likes	0.00 (CI [−0.14; 0.14])
	Wilson and Martin (2015)	Teacher	Teacher + AI	Feedback perception survey	0.09 (CI [−0.18; 0.36])
Learning dispositions	Wilson and Martin (2015)	Teacher	Teacher + AI	Problem-solving attitude	0.34 (CI [0.00; 0.68])
	Wilson and Martin (2015)	Teacher	Teacher + AI	Number of essays	0.80 (CI [0.47; 1.13])
	Darvishi et al. (2024)	Peer	Peer + AI	Metacognition	−0.02 (CI [−0.14; 0.09])

Hybrid feedback systems often aimed to leverage the strengths of both AI and human feedback. For instance, studies, such as Cheng (2022) have mapped student revisions to teacher feedback to refine AI-generated comments, thereby supporting

the revision process. Others, such as Silitonga et al. (2023), utilised conversational AI to simulate a dialogue around writing, providing students with iterative and collaborative feedback. Overall, the body of research comparing a hybrid condition with AI-only or human-only conditions is limited and scattered. However, there may be some limited evidence of a positive effect of a hybrid condition in helping students make sense of the feedback they receive from a teacher or in supporting teachers in providing feedback to students.

Discussion

This meta-analysis, based on 41 published studies (4,813 students), explored the impact of AI and human feedback on students' learning outcomes (e.g. performance), feedback perception, and dispositions (e.g. motivation, attitude). We conducted a separate meta-analysis for each learning outcome to account for the multiple outcomes and differences in study design (e.g. performance—distinguishing between task performance [single measure] and learning gains [repeated measures]; and feedback perception [single measure]). While AI holds significant potential to transform educational practices through scalable, personalised feedback, this study's findings highlight several nuances crucial for understanding and integrating AI in practical educational contexts.

The findings of this study indicate no statistically significant differences between AI-generated feedback and human-provided feedback on all the tested outcomes. However, these results should be interpreted cautiously due to the high heterogeneity observed in the data, which suggests substantial variability in the effectiveness of feedback depending on the context, AI models, and study design. For example, some AI systems provided basic corrective feedback, focusing on grammar and vocabulary errors (e.g. Chang et al., 2021; Lai, 2010), while others offered advanced features like rewriting suggestions or detailed evaluations of content and coherence (e.g. Chen & Pan, 2022; Rad et al., 2023). A critical source of variation is the underlying AI model used for generating feedback. Tools ranged from rule-based programs, like Pigai and Grammarly, to large language models, such as ChatGPT. While some newer studies employed conversational agents using LLMs (e.g. ChatGPT), most did not specify the version used (e.g. GPT-3.5 *vs.* GPT-4), limiting deeper subgroup analysis. This variation likely contributes to the heterogeneity observed in the data. As the use of specific LLMs becomes more widespread and transparently reported, future meta-analyses could examine model-specific impacts.

It is important not to interpret the lack of statistical significance as evidence of equivalence between human and AI feedback. As is the case with most studies in our datasets, they focused on 'quantifiable' outcomes and tasks that are often easy to automate or delegate to AI. In doing so, the full breadth of value that human feedback provides was not fully covered, and several essential qualities were not measured, including the ability to build confidence, foster a growth mindset, and create psychological safety within the affective and motivational domains.

Another key finding is that the current application of AI feedback is predominantly within language and writing disciplines, indicating a narrow use case of AI feedback. This focus may be attributed to the structured and rule-based nature of language,

which lends itself well to the capabilities of current AI technologies, such as grammar and style checks. For instance, AI systems like ACAwriter (Shibani et al., 2017) and EssayCritique (Mørch et al., 2017) analysed rhetorical moves and the presence or absence of subthemes in text, illustrating the tailored applications of AI feedback in writing contexts. However, this concentration raises questions about the applicability and effectiveness of AI feedback in other academic domains that may require deeper contextual understanding or more complex interaction dynamics, such as the sciences or humanities. Other fields, such as programming education, which have long relied on automated feedback (using rule-based software tests preconfigured by the teacher), were absent in our examined studies. This may be because the added benefits of AI-based feedback might be less apparent in these contexts.

Performance

The analysis revealed no significant difference between the effects of AI and human feedback on learning performance, with a small and statistically insignificant pooled effect size (Hedge's $g = 0.25$, CI [−0.11; 0.60]) and so was also the change in performance small (Hedge's $g = 0.36$) and statistically insignificant (CI [−0.37; 1.1]). This small and insignificant effect should not be viewed as a lack of efficiency but rather as evidence that AI is no less effective than human feedback and could possibly be used to automate certain aspects of feedback provision, thereby alleviating some of the workload on teachers. This is particularly significant given the scalability and resource constraints associated with human feedback, suggesting that AI could serve as a valuable complement, especially in resource-limited educational environments. Nevertheless, our studies—being mostly from writing and language learning fields—had a narrow scope, which prompts caution before drawing sweeping conclusions or generalising these findings to other fields. The variability in performance outcomes may also be partly attributable to differences in how feedback is generated and implemented; for example, Generative AI tools like ChatGPT provided immediate corrections during the writing process (e.g. Escalante et al., 2023), while rule-based AI tools like Pigai targeted essay-level performance (e.g. Zhang & Zhang, 2024). Such differences in the scope and timing of feedback likely contributed to the observed heterogeneity. Indeed, automated feedback and writing support (including AI feedback) has always proven useful to language or academic writing learners (e.g. Fleckenstein et al., 2023; Lai 2010; Ngo et al., 2024). An effect that we have not found in our study because of our broader scope beyond these fields, inclusion of many recent studies (70% in the last two years), and, more importantly, inclusion of all measures of performance, not only measures of writing efficiency.

In the three meta-analyses that we performed, there was high heterogeneity and very wide prediction intervals indicating considerable uncertainty regarding the future replication of the obtained results. In this way, researchers may obtain the full range of outcomes in future research—from highly negative to highly positive—depending on the design and contextual variables. The additional meta-analyses, which focused solely on language and writing studies, yielded similar findings, with high heterogeneity persisting in both learning performance ($I^2 = 95\%$) and feedback perception ($I^2 = 74\%$). While narrowing the scope reduced heterogeneity for feedback perception

slightly, the confidence intervals remained wide, further reinforcing the variability of results even within this subfield. These findings prompted us to ponder whether we are in a fluid moment where a new field is forming, a transitional period where further research and experimentation could refine the ways we harness AI, or if these results will hold true in subsequent applications. Only the future can tell whether innovations in AI will prove more impactful than what we see today.

These findings prompt a deeper consideration of whether AI feedback is an equivalent alternative to human feedback, as also highlighted by Escalante et al. (2023), or whether each serves distinct purposes but delivers comparable results, at least according to the current measures. For example, while AI may provide immediate and data-driven feedback, human feedback is unparalleled in understanding student emotions and providing empathetic and contextually rich responses, especially in more subjective and complex learning domains. These findings align with Hattie and Timperley's model, which emphasises that effective feedback must clarify learning goals, evaluate progress towards them, and guide future action (Feed up, Feedback, FeedForward), aspects that AI may not fully grasp without human-like contextual understanding. This finding supports the need for ongoing research into AI feedback mechanisms, as highlighted by Misiejuk et al. (2024), who advocate for a hybrid model where AI complements human judgement by providing timely feedback, while humans focus on complex and nuanced feedback that supports deep learning.

While the majority of our studies were published in the last two years, there was no statistically significant difference regarding the year of the experiment. These results show that despite technological advances, the improvement in feedback perception and learning impact delivered by AI over time has been minimal. This stagnation may suggest either a potential ceiling effect of AI capabilities in their current form, or the field of application (which was dominated by writing feedback) or a lag between the current wave of advancements in AI and their incorporation in education (Schöbel et al., 2024).

Feedback perception

Although there was no statistically significant difference between the perception of AI and human feedback, the direction of the effect size was negative. This may be because AI generates immediate and detailed feedback that often lacks the contextual nuance and personalisation that human feedback provides (Escalante et al., 2023). This underscores the potential limitations of AI in understanding the broader context of learner needs, which may hinder students' ability to self-monitor and adapt strategies, especially in situations where AI feedback lacks cues to support meta-cognitive regulation as emphasised in the self-regulated learning framework (Butler & Winne, 1995). This finding is consistent with Guo and Wang (2024) observations that while AI enhanced the quantity of feedback, its quality in terms of relevance and depth did not consistently match that provided by human educators. Moreover, advanced systems that flagged low-quality feedback and suggested improvements (e.g. Darvishi et al., 2024) or incorporated learning opportunities like vocabulary expansion (e.g. Huang, 2020) highlight how AI feedback quality can vary significantly depending on

system design and implementation. Another important factor influencing the effectiveness of feedback is the variability in the quality and delivery of human feedback. For instance, teachers and peers differ in expertise, feedback styles, and their attitudes towards students' work, all of which can moderate learning outcomes. While our subgroup analysis compared teacher and peer feedback, the primary studies rarely provided detailed data on the quality or intent of human feedback. As such, these variables could not be tested as moderators but remain important considerations for future studies.

The statistically significant effect of year observed in our analysis suggests that improvements in feedback perception over time may reflect the influence of advancements in AI technologies, particularly the emergence of large language models (LLMs). These newer AI systems, which leverage more sophisticated natural language processing techniques, may be better equipped to address contextual nuances compared to earlier rule-based systems. Aligned to Hattie and Timperly's 'Feed Forward-where to Next' dimension, AI feedback systems should advance in providing not just corrections or suggestions, but also actionable steps that are contextually aligned with students' future learning paths. It is worth mentioning that most of the studies were not blinded, meaning that students knew whether they were receiving AI- or human-generated feedback, which may have biased their perception.

Learning dispositions

Regarding learning dispositions, our study finds mixed outcomes on the impact of AI and human feedback on students' attitudes, self-regulation, motivation, and engagement. While some students appreciate the speed and accuracy of AI feedback, as Escalante et al. (2023) found, others prefer human feedback for its personalised touch. For example, Chiu et al. (2022) reported a positive effect of AI feedback on learning attitude while Lai (2010) reported a comparable negative effect size. This split preference suggests that integrating AI feedback systems in educational settings should be done thoughtfully, considering both the nature of the task and learners' personal preferences. Moreover, the nuanced differences in how feedback is perceived and its impact on learning dispositions call for a more detailed investigation into the types of learning environments and disciplines where AI might be most effective. This aligns with Wisniewski et al. (2019), who emphasised the variability in feedback effectiveness. This suggests that AI's role might be more beneficial in specific contexts or disciplines, particularly where immediate corrective feedback is valuable. As indicated by our findings, integrating AI with human feedback can enhance trust and enable more nuanced feedback, leveraging AI's efficiency and human insight. Our results suggest that there may be a positive effect of hybrid (Teacher + AI) feedback compared to AI-only or human-only feedback, which aligns with the multi-level feedback roles emphasised by Hattie and Temperley. However, there is a need for further research in this area, including the effect of the level of automation in hybrid scenarios (Molenaar, 2022). With the emergence of new AI technologies, this field of inquiry will gain more relevance as instructors strive to integrate more AI tools into their classrooms.

Implications for teaching practice

While AI offers a promising tool for addressing the scalability challenges in providing personalised feedback, it is not a panacea. Educators should consider AI as one of multiple tools in their teaching toolkit, suitable for certain types of feedback and learning scenarios, but not all. Integration of AI feedback should be approached with a clear strategy for maintaining the quality and personalisation that human feedback uniquely offers. One promising direction is the adoption of hybrid feedback systems that combine AI and human input in complementary and sequential ways. As noted by Molenaar (2022), effective hybrid intelligent systems should optimise the complementary strengths of AI (e.g. speed, scalability) and humans (e.g. empathy, contextual insight, pedagogical judgement). Moreover, hybrid systems can function sequentially (e.g. AI provides initial feedback, followed by teacher elaboration) or in parallel (e.g. students receive AI and teacher feedback simultaneously), depending on the task, learner needs, and instructional goals.

Although our study did not directly measure students' AI literacy, recent research suggests that students' ability to interpret and act on AI-generated feedback is likely influenced by their level of AI literacy. Jin et al. (2025) demonstrate, for example, how students with greater familiarity and skill in using generative AI tools perceive and utilise AI feedback more effectively. This highlights the importance of not only enhancing the quality of AI feedback systems but also supporting students in developing the competencies necessary to engage with such feedback effectively. For ed-tech developers, these findings underscore the need to design systems that are not only accurate and pedagogically aligned but also transparent and easy for students to understand and act upon, thereby supporting both effective feedback use and the development of AI literacy.

Limitations and future directions

This meta-analysis, while extensive, had several limitations that are important to consider. The primary limitation arises from the heterogeneity of the included studies. The dataset comprised 41 diverse articles, varying significantly in design, AI models used, and reported outcomes. While enriching the breadth of the analysis, this diversity complicates the synthesis of data and the drawing of broad conclusions. For example, out of the 41 included studies, 17 incorporated multiple measures per outcome, such as the number of errors, vocabulary use, and grammar. This multiplicity of measures introduces complexity in calculating and combining effect sizes, potentially leading to underestimation or overestimation of the effectiveness of AI and human feedback. To counter this problem, we used an aggregate average of all these outcomes to compute an unbiased outcome. Moreover, the variability in study designs and the inherent differences in how feedback was applied across these studies further complicate the generalisation of results. Such variability ranges from the educational levels and disciplines of the participants to the specific implementations and contexts of feedback delivery. Thus, limiting the ability to make definitive claims about the effectiveness of AI *versus* human feedback across all educational scenarios. Moreover, few studies reported the specific AI or GPT model versions used, making it challenging

to conduct subgroup analyses based on model sophistication (e.g. GPT-3.5 *vs.* GPT-4) and the particular characteristics of the human feedback provided (e.g. expertise, tone, or delivery of peer or teacher comments) which limited our ability to analyse how variation in human feedback quality or educator/peer attitudes may moderate the observed effects. In response to concerns about heterogeneity, we conducted additional meta-analyses that focused solely on studies in the field of language and writing, where the majority of the included studies were concentrated and provided specific examples of the AI models used to clarify the potential impact of the different types of AI systems. Meanwhile, the field of AI feedback in education is currently in a transitional phase. The rapid emergence of generative AI tools, such as ChatGPT has outpaced the academic publishing cycle, meaning that recent advances are not yet fully reflected in the peer-reviewed literature. As a result, the findings of this review should be interpreted as an initial synthesis of a rapidly evolving area, rather than a conclusive statement on the effectiveness of state-of-the-art AI feedback tools.

Moreover, it is worth noting that pre-post meta-analyses (Figure 4) may yield biased estimates when natural conditions influence the outcome. For example, students improve as they advance in the course between the first and second measurement points (Cuijpers, 2017). However, our studies were controlled, and we used a robust estimation method that included a conservative correlation coefficient to account for the dependence between the first and second measurement points. After all, the results of the pre-post meta-analysis did not differ significantly from those of the single treatment meta-analysis.

Conclusion

This meta-analysis contributes to the growing field of AI in education by providing evidence of its effectiveness in feedback provision compared to traditional human feedback. The findings indicate that while AI holds promise, particularly in the domain of language and writing, its effectiveness across broader academic disciplines remains uncertain due to its current predominant focus on writing evaluation. Most studies evaluated differences using quantifiable outcomes in tasks that are easy to delegate or automate using AI. Therefore, it is essential to note that a lack of statistical significance should not be interpreted as equivalence between human and AI feedback. Human feedback offers unique benefits, such as relational, mentorship, and ethical guidance that go beyond what AI can offer and most of which is still barely investigated.

The heterogeneity of the studies included and the diversity in the AI models employed suggest that future research, particularly focusing on specific AI systems or generative AI technologies, could yield more conclusive insights. However, the small and insignificant effect of AI feedback should not be viewed as a lack of efficiency, but rather as evidence of the potential of a hybrid approach where AI can augment, but not replace, the nuanced understanding and empathetic engagement that humans provide. Such hybrid systems, leveraging AI for scalability and humans for personalisation, offer a promising avenue for future feedback models in education. Finally, the diversity in methodologies, participant populations, and educational levels among the included studies, while offering broad insights, also contributes to inconsistencies and challenges in drawing generalisable conclusions. For this reason, we

position this study as an exploratory meta-analysis that reflects the current transitional moment in the field of AI-assisted feedback. Rather than aiming to draw definitive conclusions, our goal is to chart emerging patterns and limitations in the evidence base, thereby informing future meta-analyses once more consistent and model-specific studies become available.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Rogers Kaliisa http://orcid.org/0000-0001-6528-8517
Kamila Misiejuk http://orcid.org/0000-0003-0761-8703
Sonsoles López-Pernas http://orcid.org/0000-0002-9621-1392
Mohammed Saqr http://orcid.org/0000-0001-5881-3109

References

Alam, S., & Usama, M. (2023). Does E-feedback impact minimizing ESL writing errors? An experimental study. *International Journal of Emerging Technologies in Learning*, *18*(4), 156–169. https://doi.org/10.3991/ijet.v18i04.36349

Alnasser, S. M. N. (2022). EFL learners' perceptions of integrating computer-generated feedback into the writing process. *SAGE Open*, *12*(4), 1–14. https://doi.org/10.1177/21582440221123021

Azevedo, R., & Bernard, R. M. (1995). The effects of computer-presented feedback on learning from computer-based instruction: A meta-analysis.

Bearman, M., Ajjawi, R., Boud, D., Tai, J., & Dawson, P. (2023). *CRADLE suggests… assessment and genAI*. Deakin University, Centre for Research in Assessment and Digital Learning.

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

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Chang, T. S., Li, Y., Huang, H. W., & Whitfield, B. (2021, March). Exploring EFL students' writing performance and their acceptance of AI-based automated writing feedback. In *Proceedings of the 2021 2nd International Conference on Education Development and Studies* (pp. 31–35). https://doi.org/10.1145/3459043.3459065

Chen, H. M., & Pan, J. (2022). Computer or human: A comparative study of automated writing evaluation and teacher feedback in the EFL context. *Asian-Pacific Journal of Second and Foreign Language Education*, *7*(1), 1–18. https://doi.org/10.1186/s40862-022-00171-4

Cheng, G. (2022). Exploring the effects of automated tracking of student responses to teacher feedback in draft revision: Evidence from an undergraduate EFL writing course. *Interactive Learning Environments*, *30*(2), 353–375. https://doi.org/10.1080/10494820.2019.1655769

Cheng, G., Law, E., & Wong, T.-L. (2017). Investigating effects of automated feedback on EFL students' critical thinking skills. In *Proceedings of the IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (pp. 270–276). IEEE. https://doi.org/10.1109/TALE.2016.7851798

Chiu, M. C., Hwang, G. J., Hsia, L. H., & Shyu, F. M. (2022). Artificial intelligence-supported art education: A deep learning-based system for promoting university students' artwork appreciation and painting outcomes. *Interactive Learning Environments*, *32*(1), 1–26. https://doi.org/10.1080/10494820.2022.2100426

Cuijpers, P. (2017). Four decades of outcome research on psychotherapies for adult depression: An overview of a series of meta-analyses. *Canadian Psychology*, *58*(1), 7–19. https://doi.org/10.1037/cap0000096

Darvishi, A., Khosravi, H., Abdi, S., Kitto, K., & Gašević, D. (2022). Incorporating training, self-monitoring and AI-based feedback to improve students' academic writing. In *Proceedings of the ACM Conference on Learning@Scale* (pp. 215–225). ACM. https://doi.org/10.1145/3491140.3528265

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*, *210*, 104967. https://doi.org/10.1016/j.compedu.2023.104967

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint arXiv:1810.04805.

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*, *29*(11), 14151–14203. https://doi.org/10.1007/s10639-023-12402-3

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

Ebadi, S., Fathi, J., & Rashidi, N. (2023). Exploring the impact of automated feedback on EFL learners' writing performance and engagement. *Interactive Learning Environments*, *31*(4), 555–572. https://doi.org/10.1080/10494820.2021.1954039

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 57. https://doi.org/10.1186/s41239-023-00425-2

Fan, N. (2023). Exploring the effects of automated written corrective feedback on EFL students' writing quality: A mixed-methods study. *Sage Open*, *13*(2), 21582440231181296. https://doi.org/10.1177/21582440231181296

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, *6*, 1162454. https://doi.org/10.3389/frai.2023.116245

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, *11*(12), 5467. https://doi.org/10.3390/app11125467

Green, S., & Higgins, J. P. T. (2011). Cochrane handbook for systematic reviews of interventions.

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Hamer, J., Purchase, H., Luxton-Reilly, A., & Denny, P. (2015). A comparison of peer and tutor feedback. *Assessment & Evaluation in Higher Education*, *40*(1), 151–164. https://doi.org/10.1080/02602938.2014.893418

Hansen, C., Steinmetz, H., & Block, J. (2022). How to conduct a meta-analysis in eight steps: A practical guide. *Management Review Quarterly*, *72*(1), 1–19. https://doi.org/10.1007/s11301-021-00247-4

Hassanzadeh, M., & Fotoohnejad, S. (2021). Implementing an automated feedback program for a Foreign Language writing course: A learner-centric study: Implementing an AWE tool in a L2 class. *Journal of Computer Assisted Learning*, *37*(5), 1494–1507. https://doi.org/10.1111/jcal.12587

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

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

Hu, X., Li, Y., & Zheng, Y. (2023). Chatbot-assisted writing feedback: Investigating EFL students' perceptions and outcomes. *Interactive Learning Environments*, *31*(2), 220–237. https://doi.org/10.1080/10494820.2023.2208170

Huang, L. (2020). From feedback to revision: How can collaborative peer feedback activities improve students' writing? In *Proceedings of the International Conference on Educational Innovation through Technology (EITT)* (pp. 52–56). IEEE. https://doi.org/10.1109/EITT53287.2021.00052

Iraj, H., Fudge, A., Khan, H., Faulkner, M., Pardo, A., & Kovanović, V. (2021). Narrowing the feedback gap: Examining student engagement with personalized and actionable feedback messages. *Journal of Learning Analytics*, *8*(3), 101–116. https://doi.org/10.18608/jla.2021.7184

Jin, Y., Yang, K., Yan, L., Echeverria, V., Zhao, L., Alfredo, R., & Martinez-Maldonado, R. (2025, March). Chatting with a learning analytics dashboard: The role of generative AI literacy on learner interaction with conventional and scaffolding chatbots. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 579–590).

Johnson, N., & Phillips, M. (2018). Rayyan for systematic reviews. *Journal of Electronic Resources Librarianship*, *30*(1), 46–48. https://doi.org/10.1080/1941126X.2018.1444339

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

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Knight, S., Shibani, A., Abel, S., Gibson, A., & Ryan, P. (2020). AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research*, *12*(1), 141–186. https://doi.org/10.17239/jowr-2020.12.01.06

Lai, Y. H. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology*, *41*(3), 432–454. https://doi.org/10.1111/j.1467-8535.2009.00959.x

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004

Lin, J., Song, J., & Sun, L. (2020, June). The application of artificial intelligence video feedback system in tennis teaching in colleges and universities. In *2020 International Conference on Artificial Intelligence and Education (ICAIE)* (pp. 28–31). IEEE. https://doi.org/10.1109/ICAIE50891.2020.00014

Liu, L. (2022, December). Application of automated writing evaluation (AWE) system based on intelligent technology. In *2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICMNWC56175.2022.10031924

Liu, M., Li, Y., Xu, W., & Liu, L. (2017). Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies*, *10*(4), 502–513. https://doi.org/10.1109/TLT.2016.2612659

Lu, M., Deng, Q., & Yang, M. (2020). EFL writing assessment: Peer assessment vs. automated writing evaluation. In *Lecture Notes in Computer Science* (Vol. 12063, pp. 25–35). https://doi.org/10.1007/978-3-030-38778-5

Misiejuk, K., Kaliisa, R., & Scianna, J. (2024). Augmenting assessment with AI coding of online student discourse: A question of reliability. *Computers and Education: Artificial Intelligence*, *6*, 100216. https://doi.org/10.1016/j.caeai.2024.100216

Mohammadi, M., Zarrabi, M., & Kamali, Z. (2023). Automated writing evaluation in EFL classrooms: Impacts on learners' writing quality and perceptions. *International Journal of Language Testing*, *13*(1), 45–61.

Mohsen, M. A., & Alshahrani, A. (2019). The effectiveness of using a hybrid mode of automated writing evaluation and teacher feedback to improve EFL students' writing. *Teaching English with Technology*, *19*(3), 122–139.

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, *57*(4), 632–645. https://doi.org/10.1111/ejed.12527

Mørch, A. I., Engeness, I., Cheng, V. C., Cheung, W. M., & Wong, K. C. (2017). EssayCritic: Writing feedback using automated and peer assessments. *Educational Technology and Society*, *20*(3), 107–120.

Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments*, *32*(2), 727–744. https://doi.org/10.1080/10494820.2022.2096642

Ochoa, X., & Domínguez, F. (2020). Automatic presentation feedback: Improving students' oral skills in higher education. *British Journal of Educational Technology*, *51*(5), 1804–1817. https://doi.org/10.1111/bjet.12950

Olkin, I., Dahabreh, I. J., & Trikalinos, T. A. (2012). GOSH—A graphical display of study heterogeneity. *Research Synthesis Methods*, *3*(3), 214–223. https://doi.org/10.1002/jrsm.1053

OpenAI. (2022, November 30). *Introducing ChatGPT*. OpenAI. https://openai.com/index/chatgpt/

Ouyang, F., Wu, M., Zheng, L., Zhang, L., & Jiao, P. (2023). Integrating learning analytics and AI feedback in undergraduate writing. *International Journal of Educational Technology in Higher Education*, *20*(1), 4. https://doi.org/10.1186/s41239-022-00372-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ (Clinical Research ed.)*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pirttinen, N., & Leinonen, J. (2022). Can students review their peers? Comparison of peer and instructor reviews. In *Proceedings of the 27th ACM Conference on Innovation and Technology in Computer Science Education* (Vol. 1, pp. 12–18).

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

Rad, H. S., Alipour, R., & Jafarpour, A. (2023). AI feedback using AI Tune: A study of second language learners. *Interactive Learning Environments*, *32*(9), 5020–5040. https://doi.org/10.1080/10494820.2023.2208170

Rosen, Y. (2015). Computer-based assessment of collaborative problem solving: Exploring the feasibility of human-to-agent approach. *International Journal of Artificial Intelligence in Education*, *25*(3), 380–406. https://doi.org/10.1007/s40593-015-0042-3

Ruegg, R. (2015). Differences in the uptake of peer and teacher feedback. *RELC Journal*, *46*(2), 131–145. https://doi.org/10.1177/0033688214562799

Ruwe, T., & Mayweg-Paus, E. (2023). AI-supported argumentative writing feedback: Impacts on students' revision behavior. *Computers and Education: Artificial Intelligence*, *5*, 100189. https://doi.org/10.1016/j.caeai.2023.100189

Schöbel, S., Schmitt, A., Benner, D., Saqr, M., Janson, A., & Leimeister, J. M. (2024). Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers. *Information Systems Frontiers*, *26*(2), 729–754. https://doi.org/10.1007/s10796-023-10375-9

Shang, H.-F. (2022). Exploring online peer feedback and automated corrective feedback in EFL writing. *Interactive Learning Environments*, *30*(1), 4–16. https://doi.org/10.1080/10494820.2019.1629601

Shibani, A., Knight, S., & Buckingham Shum, S. (2017). Design and implementation of a pedagogic intervention using automated feedback in academic writing. In *Proceedings of the International Conference on Computers in Education (ICCE)* (pp. 65–74).

Silitonga, L. M., Hawanti, S., & Aziez, F. (2023). The impact of AI chatbot-based learning on students' writing outcomes. In *Lecture Notes in Computer Science* (Vol. 13991, pp. 678–690). https://doi.org/10.1007/978-3-031-40113-8_53

Sun, B., & Fan, T. (2022). The effects of an AWE-aided assessment approach on EFL business English writing performance and perceptions. *Studies in Educational Evaluation*, *72*, 101123. https://doi.org/10.1016/j.stueduc.2021.101123

Taskiran, A., & Goksel, N. (2022). Automated feedback and teacher feedback: Writing skills in distance education. *Turkish Online Journal of Distance Education*, *23*(2), 45–60.

Van der Pol, J., Van den Berg, B. A. M., Admiraal, W. F., & Simons, P. R. J. (2008). The nature, reception, and use of online peer feedback in higher education. *Computers & Education*, *51*(4), 1804–1817. https://doi.org/10.1016/j.compedu.2008.06.001

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P. T., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*(1), 55–79. https://doi.org/10.1002/jrsm.1164

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Waer, H. (2023). Automated writing evaluation and teacher feedback: A comparative study in EFL contexts. *Language Teaching Research*, *27*(3), 320–339. https://doi.org/10.1177/13621688211041165

Wang, J. (2019). A comparative study on the washback effects of teacher feedback plus intelligent feedback versus teacher feedback on English writing teaching in higher vocational college. *Theory and Practice in Language Studies*, *9*(12), 1555–1562. https://doi.org/10.17507/tpls.0912.12

Wang, Y. J., Shang, H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*, *26*(3), 234–257. https://doi.org/10.1080/09588221.2012.655300

Wang, Z., & Han, F. (2022). The effects of teacher feedback and automated writing evaluation on EFL students' writing. *Frontiers in Psychology*, *13*, 909802. https://doi.org/10.3389/fpsyg.2022.909802

Weitekamp, D., Harpstead, E., & Koedinger, K. R. (2020, April). An interaction design for machine teaching to develop AI tutors. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). https://doi.org/10.1145/3313831.3376226

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on student writing. *Computers & Education*, *92–93*, 82–95. https://doi.org/10.1016/j.compedu.2016.05.004

Wilson, J., & Martin, T. (2015). Using PEGWriting® to support the writing motivation and performance of middle school students. *Workshop on Innovative Use of NLP for Building Educational Applications (BEA)* (pp. 45–52). ACL.

Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology*, *60*, 101826. https://doi.org/10.1016/j.cedpsych.2019.101826

Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Automated writing feedback versus teacher feedback. *Computers & Education*, *161*, 104059. https://doi.org/10.1016/j.compedu.2020.104059

Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, *61*(4), 875–900. https://doi.org/10.1016/j.asej.2014.09.007

Zhang, J., & Zhang, L. J. (2024). The effect of feedback on metacognitive strategy use in EFL writing. *Computer Assisted Language Learning*, *37*(5–6), 1198–1223. https://doi.org/10.1080/09588221.2022.2069822

Appendix A

Table A1 presents all the studies included in the meta-analysis and outcomes measured that reported enough data for meta-analysis. Tables A2–A4 present all the studies that each of the three main meta-analyses presented in this study—task performance, learning gains, and feed-

Table A1. Studies included in the meta-analysis and outcomes measured that reported enough data for meta-analysis.

Reference	Educational level	Discipline	Experimental	Control	P	FPQ	LD
Alam and Usama (2023)	Higher education	Language/ writing	AI (Grammarly)	Peer	●		
Alnasser (2022)	Higher education	Language/ writing	AI	Peer		●	
Chang et al. (2021)	Higher education	Language/ writing	AI (Grammarly)	Peer	●		
Chen and Pan (2022)	Higher education	Language/ writing	AI (iWrite)	Peer	●		
Cheng et al. (2017)	Higher education	Language/ writing	AI	Teacher	●		
Cheng (2022)	Higher education	Language/ writing	Teacher + AI	Teacher	●		
Chiu et al. (2022)	Higher education	STEAM	AI (DL-AL)	Teacher		●	●
Darvishi et al. (2022)	Higher education	STEAM	Peer + AI (BERT)	Peer		●	
Darvishi et al. (2024)	Higher education	STEAM	Peer + AI (BERT)	Peer		●	●
Ebadi et al. (2023)	Higher education	Language/ writing	AI (Grammarly)/ hybrid	Teacher/hybrid	●		
Escalante et al. (2023)	Higher education	Language/ writing	AI (ChatGPT)	Teacher	●	●	
Fan (2023)	Higher education	Language/ writing	Teacher + AI (Grammarly)	Teacher	●		
Hassanzadeh and Fotoohnejad (2021)	Higher education	Language/ writing	AI (Criterion)	Teacher	●		
Hu et al. (2023)	Higher education	STEAM	AI (DRL Agent)	Teacher	●		
Huang (2020)	Higher education	Language/ writing	Peer + AI (Pigai)	Peer	●		
Lai (2010)	Higher education	Language/ writing	AI (MY Access)	Peer	●	●	●
Lin et al. (2020)	Higher education	Sports	Teacher + AI	Teacher	●		
Liu (2022)	Higher education	Language/ writing	AI (iWrite)	Teacher	●		
Liu et al. (2017)	Higher education	Language/ writing	AI (SAM)	Teacher	●		
Lu et al. (2020)	Higher education	Language/ writing	AI (Pigai)	Peer		●	
Mohammadi et al. (2023)	Higher education	Language/ writing	Teacher + AI (Virtual Writing Tutor)	Teacher	●		
Mohsen and Alshahrani (2019)	Higher education	Language/ writing	AI (MY Access)	Hybrid (Teacher + AI)	●		
Mørch et al. (2017)	Higher education	Language/ writing	AI (EssayCritic)	Peer	●		
Ochoa and Domínguez (2020)	Higher education	Language/ writing	AI (RAP)	Teacher	●		
Ouyang et al. (2023)	Higher education	STEAM	AI (Unspecified)	Teacher	●		●

(*Continued*)

Table A1. Continued.

Reference	Educational level	Discipline	Experimental	Control	P	FPQ	LD
Rad et al. (2023)	Higher education	Language/ writing	AI (Wordtune)	Teacher + Peer	●		●
Rosen (2015)	K-12	STEAM	AI (Unspecified)	Peer	●		●
Ruwe and Mayweg-Paus (2023)	Higher education	Language/ writing	AI (NLP)	Peer/teacher		●	●
Shang (2022)	Higher education	Language/ writing	AI (Pigai)	Peer	●	●	
Shibani et al. (2017)	Higher education	Social science	AI (Academic Writing Analytics)	Teacher		●	
Silitonga et al. (2023)	Higher education	Language/ writing	AI (ChatGPT)	Teacher			●
Sun and Fan (2022)	Higher education	Language/ writing	AI (Pigai)	Teacher	●		●
Taskiran and Goksel (2022)	Open	Language/ writing	AI (Write & Improve)	Teacher	●		
Waer (2023)	Higher education	Language/ writing	AI (Write & Improve)	Teacher	●		●
Wang (2019)	Higher education	Language/ writing	Teacher + AI (Pigai)	Teacher	●		
Wang et al. (2013)	Higher education	Language/ writing	AI (CorrectEnglish)	Teacher	●		
Wang and Han (2022)	Higher education	Language/ writing	AI (Pigai)	Teacher	●	●	
Wilson and Martin (2015)	K-12	Language/ writing	Teacher + AI (PEG Writing)	Teacher		●	●
Wilson and Czik (2016)	K-12	Language/ writing	Teacher + AI (PEG Writing)	Teacher	●		
Xu et al. (2021)	K-12	Language/ writing	AI (Unspecified)	An adult	●		
Zhang & Zhang (2024)	Higher education	Language/ writing	AI (Pigai)	Peer/Teacher	●		●

P: performance; FPFQ: feedback perception quality; LD: learning dispositions.

Table A2. Influence, sensitivity, and bias analysis for the performance meta-analysis.

Analysis	Hedge's *g*	95%CI	I^2	95%CI
Main analysis	0.25	−0.11–0.60	75.0%	56.0–85.8%
Simple outliers removed[a]	0.19	0.01–0.36	25.4%	0.0–64.0%
Leave-one-out influential cases removed[b]	0.12	−0.13–0.36	57.1%	16.0–78.1%
Trim-and-fill method[c]	0.09	−0.31–0.48	81.4%	69.9–88.5%
GOSH analysis outliers removed[d]	0.29	−0.10–0.69	75.3%	55.3–86.3%

[a]Removed as outliers: Lai (2010) and Wang et al. (2013).
[b]Removed influential studies: Wang et al. (2013).
[c]With 2 added studies as fill: Ouyang et al. (2023) and Wang et al. (2013).
[d]Removed 1 studies identified by DBScan clustering: Shang (2022).

Table A3. Influence, sensitivity, and bias analysis for the learning gains meta-analysis.

Analysis	Hedge's *g*	95%CI	I^2	95%CI
Main analysis	0.36	−0.37–1.10	94.0%	91.6–95.9%
Simple outliers removed[a]	0.28	−0.18–0.73	92.1%	87.9–94.9%
Leave-one-out influential cases removed[b]	0.45	−0.35–1.24	93.9%	91.1–95.8%
Trim-and-fill method[c]	0.36	−0.37–1.10	94.0%	91.6–95.9%
GOSH analysis outliers removed[d]	0.13	−0.36–0.63	93.4%	90.3–95.5%

[a]Removed: Alam and Usama (2023) and Hassanzadeh and Fotoohnejad (2021).
[b]Removed influential studies: Wang and Han (2022).
[c]With 0 added studies.
[d]Removed 1 studies identified by DBScan clustering: Hassanzadeh and Fotoohnejad (2021).

Table A4. Influence, sensitivity, and bias analysis for the feedback perception meta-analysis.

Analysis	Hedge's *g*	95%CI	I^2	95%CI
Main analysis	−0.20	−0.67–0.27	84.7%	72.7–91.4%
Simple outliers removed[a]	−0.35	−0.77–0.06	75.4%	50.6–87.8%
Leave-one-out influential cases removed[b]	−0.20	−0.67–0.27	84.7%	72.7–91.4%
Trim-and-fill method[c]	−0.17	−0.60–0.27	84.7%	72.7–91.4%
GOSH analysis outliers removed[d]	−0.11	−0.60–0.39	79.8%	60.8–89.6%

[a]Removed as outliers: Chiu et al. (2022).
[b]Removed influential studies: None.
[c]With 0 added studies.
[d]Removed 1 studies identified by DBScan clustering: Shibani et al. (2017).

back perception—computed after applying several techniques related to outlier removal (through sensitivity or influence analysis), or bias correction. For each of the three meta-analyses, we present:

1. **Main analysis:** The original meta-analysis presented in the main body of the paper for reference
2. **Simple outliers removed:** A meta-analysis performed in all but those studies where either limit of the confidence interval lay outside the confidence interval of all studies.
3. **Leave-one-out influential cases removed**: A meta-analysis performed after removing the outliers identified through the leave-one-out sensitivity analysis, which are studies that significantly skew the effect size (Viechtbauer & Cheung, 2010).
4. **Trim-and-fill method:** A meta-analysis performed after detecting and adjusting for asymmetries in funnel plots that might indicate missing studies (Duval & Tweedie, 2000).
5. **GOSH analysis outliers removed:** A meta-analysis conducted after identifying and excluding outliers through the GOSH analysis (Olkin et al., 2012).