Educational Psychology Review (2026) 38:13
https://doi.org/10.1007/s10648-026-10117-8

META-ANALYSIS

A Meta-Analysis of the Impact of Feedback Timing on Learning Outcomes in Computer-Assisted Learning

Erva Nihan Kandemir[1] · **Emanuele Esposito**[1] · **Lilas Gurgand**[1] · **Franck Ramus**[1]

Received: 5 April 2025 / Accepted: 1 January 2026
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

Despite decades of research, the impact of feedback timing on learning outcomes remains debated. The most recent meta-analysis directly comparing immediate and delayed feedback was published in 1988, preceding the transformation of computer-assisted learning environments. This study reevaluates the effect of feedback timing on learning outcomes within computer-assisted learning environments. Three key aspects were examined: (a) the overall effectiveness of immediate versus delayed feedback, (b) the influence of varying definitions of feedback delay on learning outcomes, and (c) study characteristics and feedback features that may account for prior inconsistencies. A systematic meta-analysis was conducted on 51 studies published from 1988 to 2024, analyzing 160 effect sizes through meta-regression with robust variance estimation. Across the included studies, delays implemented in delayed feedback conditions ranged from 1 second to 7 days in time-based designs and from 1 to 60 intervening items in item-based designs. Results indicate that feedback timing does not significantly influence learning outcomes on average ($g = 0.03$, 95% CI $[-0.08, 0.13]$, $p = .61$). Substantial heterogeneity between studies suggests the effect is moderated by key factors. Moderator analyses reveal that educational level, learning domain, and response time constraints significantly influence feedback timing effectiveness, providing a partial explanation for inconsistencies across studies. Yet, partial confounding of moderators limits the ability to fully disentangle their individual contributions. These findings suggest that while feedback timing alone is not a decisive factor, its impact is context-dependent. Although approximately half of the effects involved multi-hour or $\geq$ 10-item delays, there were relatively few studies involving delays of one day or longer. As a result, the reviewed studies provide limited evidence concerning the common multi-day delays typical of many classroom settings.

Keywords Learning · Feedback timing · Meta-analysis · Optimal feedback

Extended author information available on the last page of the article

Published online: 26 January 2026

Introduction

Feedback refers to information provided regarding learners' performance that is aimed at assisting them in confirming, expanding, or modifying their stored knowledge (Butler & Winne, 1995; Hattie & Timperley, 2007; Sadler, 1989). The critical impact of feedback on education is well-documented through extensive meta-analyses (Hattie & Clarke, 2018; Hattie & Timperley, 2007; Wisniewski et al., 2020).

With the rapid integration of technology into education, digital learning environments are now offering advanced feedback mechanisms that can enhance learning outcomes more effectively than traditional methods, although the degree of this impact varies based on the specific characteristics of the feedback (Azevedo & Bernard, 1995; Cai et al., 2023; Mohamed & Adnan, 2020; Mohsen, 2022; Van der Kleij et al., 2015). One notable feature of digital learning environments is the systematicity and precision with which they can deliver feedback. Unlike traditional educational methods that relied mostly on teacher correction, digital learning environments can provide feedback almost immediately after a response is recorded. This capability has significantly advanced the scope of research into feedback timing.

Research on the impact of feedback timing has a long history. This history was summarized in the seminal meta-analysis by Kulik and Kulik (1988), which investigated whether receiving feedback immediately or after a delay was more effective for learning. Although few studies at the time used digital tools, the findings indicated that immediate feedback generally had small to moderate positive effects, compared to delayed feedback. However, subsequent research has provided mixed results, leaving the role of feedback timing in digital learning environments unclear. To the best of our knowledge, no meta-analysis has specifically focused on studies that directly compared immediate and delayed feedback since Kulik and Kulik (1988). This study seeks to address this gap through a comprehensive meta-analytic review of empirical research.

Theoretical Perspectives on Feedback Timing

In their review, Shute (2008) classified feedback timing in educational settings into two categories: immediate and delayed. Immediate feedback is defined as feedback given right after a learner completes a task or a problem-solving step. Delayed feedback, on the other hand, is provided after an extended period, ranging from several hours to several days, or even up to a week following task completion. This categorization naturally led to a debate about which timing was more beneficial for learning, prompting the development of multiple theoretical explanations to account for the observed effects (see Butler and Woodward 2018 for a detailed review).

Early support for immediate feedback is rooted in behaviorist theory (Hull, 1952; Saltzman, 1951; Skinner, 1965), which suggests that feedback (or reinforcement) must be immediate to effectively influence learning (or conditioning). Additionally, cognitive load theory (Sweller, 2011) also supports immediate feedback by indicating that working memory's limited capacity is best managed when learners receive detailed feedback in digestible amounts, thus avoiding cognitive overload.

On the other hand, several frameworks support the benefits of delayed feedback. One of the earliest and most influential is the interference-perseveration hypothesis (Kulhavy & Anderson, 1972), which argues that delayed feedback is more effective for correcting errors because the delay allows initially incorrect responses to fade from working memory, therefore minimizing confusion and enhancing the acquisition of the correct response. This theory assumes that errors hinder learning. However, more recent evidence challenges this view by showing that learners often remember their errors even after a delay, and that identification of the initial answer is highest when the responses were initially incorrect but later corrected on a final test (Peeck et al., 1985; Peeck & Tillema, 1979), suggesting that generating errors can actually facilitate learning (Butler et al., 2011; Metcalfe & Xu, 2018). Moreover, although the interference-perseveration hypothesis assumes that erroneous responses must be forgotten for feedback to be effective, numerous studies have shown that delayed feedback can still enhance learning even when it explicitly reintroduces learners to their original errors during the feedback phase (e.g., Surber and Anderson 1975; Butler et al. 2007; Butler and Roediger 2008).

These findings highlight that the interference-perseveration may be limited even when considering only initially incorrect responses, and that its assumptions are increasingly contested. Moreover, any explanation for the benefits of delayed feedback must also account for its positive effects on initially correct responses (Butler et al., 2007; Butler & Roediger, 2008; Kulhavy & Anderson, 1972).

One such explanation is offered by the dual-trace hypothesis (Clariana et al., 2000; Kulik & Kulik, 1988), which proposes that delayed feedback fosters learning by offering two distinct encoding opportunities: one at the time of the initial response and another during feedback, effectively reinforcing correct responses and correcting errors. Another explanation, the attention-based account (Kulhavy & Anderson, 1972; Phye & Andre, 1989), suggests that delayed feedback is more effective because learners are more likely to pay attention to and engage with feedback that is separated from the learning event, giving them more time to process and understand the feedback.

More recent theories propose a memory-based explanation for the benefits of delayed feedback. One prominent perspective emphasizes the role of distributed practice. According to this view, delayed feedback does not enhance learning because of the delay itself, but because it creates an additional spaced-retrieval opportunity. Research on the spacing effect shows that distributing retrieval attempts over time strengthens memory retention more effectively than massed practice (Carpenter et al., 2012; Cepeda et al., 2008; Latimier et al., 2021). Thus, improvements following delayed feedback may not reflect a true effect of feedback timing per se, but rather the benefits of spaced retrieval and memory consolidation, particularly for initially correct responses (Butler et al., 2007; Metcalfe et al., 2009).

Complementing these behavioral theories, neurocognitive evidence suggests that different timing conditions may recruit distinct memory systems. Studies show that immediate feedback engages reward-sensitive regions, while delayed feedback involves structures involved in relational memory formation (Foerde & Shohamy, 2011).

Previous Meta-analyses and Reviews

Immediate feedback is often preferred over delayed feedback by educators and particularly by students (Lefevre & Cox, 2017; Mullet et al., 2014; Van der Kleij et al., 2012), under the assumption that it facilitates better learning outcomes. This preference has influenced the design of many online courses, learning platforms, and applications, which are structured to provide immediate feedback (e.g., Kahoot!, Duolingo, Quizlet, ALEKS). However, a detailed examination of the literature reveals mixed results.

The meta-analysis by Kulik and Kulik (1988), included 53 studies from 1966 to 1988 specifically comparing these two feedback timings. They found that immediate feedback generally outperformed delayed feedback in classroom studies, while in laboratory settings, delayed feedback was more effective. The meta-analysis also explored to what extent the definition of the delay, either after each item or after the entire test, affected outcomes, and revealed nuanced differences in the effectiveness of these two definitions.

Other meta-analyses and reviews have more generally investigated the impact of feedback on learning outcomes, using feedback timing as a moderator. For instance, Azevedo and Bernard (1995) assessed the influence of computerized feedback on learning by synthesizing 22 studies published until 1992. They found that immediate feedback significantly enhanced students' academic achievement compared to delayed feedback. A more recent analysis by Van der Kleij et al. (2015) examined feedback effects in computerized settings across 40 studies from 1968 to 2012. They categorized feedback timing into immediate and delayed, defining delayed feedback as any feedback "not delivered immediately after completing each item". Their findings indicated that immediate feedback was particularly effective for lower-order learning tasks, whereas delayed feedback seemed more suitable for higher-order tasks, though they reported no significant interaction between these moderators. Swart et al. (2019) conducted a meta-analysis of 60 studies from 1962 to 2016, focusing on the feedback effect on reading performance. Their moderator analysis revealed that feedback was most effective when provided immediately after a learning session rather than during the session itself. Furthermore, in their review, Jaehnig and Miller (2007) examined 33 studies from 1964 to 2004, with various feedback types and timings. They found no significant differences in learning outcomes between immediate and delayed feedback.

Finally, a very recent review Xu and Zeng (2023) investigated 20 empirical studies on second language learning from 2006 to 2021, which directly compared immediate and delayed feedback. The results suggested that immediate feedback was often more effective or equally effective compared to delayed feedback, with variations likely due to differences in communication modalities, feedback explicitness, and the specific timing of delays. Thus, the role of feedback timing may not be constant across studies but may depend on specific learning contexts and study characteristics, such as educational level, learning domain, task design (experimental or curriculum-based), retention interval, post-test timing constraints, and the way delay is defined.

Possible Moderators of Feedback Timing Effects

Previous review studies on feedback effectiveness have primarily focused on core characteristics such as timing and type (Van der Kleij et al., 2012, 2015). More recently, attention has expanded to include additional dimensions like presentation format (Donkin et al., 2019) and adaptability to learner needs (Janelli & Lipnevich, 2021; Lim et al., 2021). Research also suggests that factors such as learners' prior knowledge, learning environment, and outcome type can influence feedback effectiveness in digital learning contexts (Swart et al., 2019; Van der Kleij et al., 2015).

Drawing from this literature, we identified key moderators that may explain inconsistent findings regarding feedback timing. These fall into two categories: (1) experimental design features—such as feedback type, how delay is defined, the actual time gap between feedback conditions, and task context; and (2) learner-related factors—including educational level, subject domain, and task complexity.

Feedback Type

Feedback type is a key characteristic of educational feedback and has been widely examined in meta-analyses and reviews. Shute (2008) categorized feedback into two types: simple and elaborated feedback. Simple feedback, comprising knowledge of results (KR) and knowledge of correct response (KCR), provides basic right-wrong information and, in the case of KCR, the correct answer. Elaborated feedback (EF), however, not only indicates correctness but also provides detailed explanations and guidance for improvement.

While Brummer et al. (2024) reported no significant effects of feedback type on learning, Cai et al. (2023) found that certain types of EF are significantly more effective than simpler forms. These findings align with Van der Kleij et al. (2015) and Mertens et al. (2022) who noted a consistent trend across studies where richer feedback content more effectively enhanced learning outcomes. Moreover, Swart et al. (2019) explored the interaction between feedback type and timing and revealed that EF and KCR are more effective when provided after, rather than during reading.

Individual studies examining the interaction between feedback type and timing also offer diverse insights. For example, Smits et al. (2008); Corral et al. (2021) found no significant interaction when comparing EF with KCR. However, Taxipulati and Lu (2021) identified a significant interaction, highlighting that learners responded more attentively to adaptive feedback when it was provided immediately rather than delayed. Furthermore, Roper (1977) suggested combining immediate KR with delayed EF, suggesting that quick correctness verification followed by a period for reflection best enhances learning.

This variability underscores the necessity of considering both feedback type and timing in educational research to better understand the inconsistent findings in prior feedback timing research.

Definition of Feedback Delay and Delay Difference Between Immediate and Delayed Feedback

Immediate feedback typically occurs right after a student responds to a question. The definition of delayed feedback, however, can span from a few hours to a week following task completion (Dempsey & Wager, 1988; Shute, 2008). With advancements in technology, particularly in computer-based environments, Van der Kleij et al. (2015) recommends considering any feedback that is not immediate in such contexts as delayed. However, as delivery methods diversified, two distinct ways to define feedback delay have emerged. Some studies in computer-based environments define delay by the number of intervening items before feedback is given, ranging from a few items to the end of an entire task, while others define it by time, providing feedback after a single item or after completing the entire task, but with delays spanning several seconds, hours, or even a week.

This variability clearly shows that the definition of feedback delay may be an important contributor to inconsistent findings in research on the impacts of immediate versus delayed feedback (Kulik & Kulik, 1988; Mory, 2013; Xu & Zeng, 2023). This suggests that merely categorizing feedback as immediate or delayed, as seen in previous meta-analyses, may not sufficiently capture the nuances affecting learning outcomes. Instead, ensuring that studies are comparable requires aligning them on the same delay scale—whether in terms of items or time. Then, the precise delay difference between feedback defined as immediate and as delayed in each study can be used as moderator, allowing for a consistent and reliable comparison across different studies.

Learning Task Context

The learning process can differ significantly depending on whether the task is designed purely for research purposes, or if it is integrated into a real curriculum, directly benefiting learners' educational or professional goals. Although not explicitly categorizing these two contexts, Kulik and Kulik (1988) explored the effects of feedback timing in both classroom and laboratory settings. They concluded that in classroom studies, immediate feedback tends to enhance learning more effectively, whereas in laboratory studies, delayed feedback often proves superior. This disparity may be attributed to differences in motivation, cognitive load, and relevance to real-world applications, which are also relevant to the distinction between experimental and curriculum-based learning tasks. Therefore, the context of the learning task should be thoroughly investigated as a potential moderator of the feedback timing effect.

Education Level

The impact of immediate versus delayed feedback on learning may differ by educational level, as there are cognitive and engagement differences between students at different levels (Swart et al., 2019). While a meta-analysis by Van der Kleij et al. (2015) found that computerized feedback was more effective for college students than

for those in primary and middle school, the role of educational level as a moderator of feedback timing has not been extensively tested since Kulik and Kulik (1988)'s findings, which found no significant effect. Considering the advancements in computerized education and additional research conducted since 1988, education level may be a critical moderator in the effectiveness of feedback timing across studies.

Learning Domain

The optimal timing for correcting erroneous knowledge may vary across different learning domains due to their unique characteristics. Research by Van der Kleij et al. (2015) suggested that computer-based feedback significantly enhanced math performance but had minimal impact on language learning. However, this finding was influenced by the predominance of elaborated feedback in math in the studies reviewed and requires cautious interpretation. Furthermore, the results of Mertens et al. (2022) did not show larger effect sizes in mathematics compared to other domains but did reveal that the effectiveness of specific feedback types varied by learning domain. These observations highlight the need to examine the role of the learning domain as a moderator of feedback timing effects.

Training Task Complexity

Training task complexity within a learning domain may also influence the effectiveness of feedback timing. Lower-order tasks, involving knowledge, comprehension, recall, understanding, and reproduction, differ from higher-order tasks, which require application, analysis, synthesis, evaluation, and strategic thinking (Bloom et al., 1964). According to Mason and Bruning (2001), immediate feedback is more beneficial for lower-achieving students regardless of task level, whereas higher-achieving students benefit more from delayed feedback, particularly with higher-order tasks. Shute (2008) concurs, suggesting that immediate feedback is optimal for lower-order learning outcomes, while higher-order outcomes benefit from a delay. Similarly, Van der Kleij et al. (2015) observed that immediate feedback enhanced lower-order task performance, but found no significant interaction between learning outcome levels and feedback timing. Therefore, training task complexity should be considered a potential moderator of the effectiveness of immediate versus delayed feedback.

Limitations of Existing Meta-analyses on Feedback Timing

Although previous meta-analyses shed light on the effects of feedback timing on learning outcomes, they present limitations, especially in digital learning environments. Firstly, the variability in the definitions of "immediate" and "delayed" feedback across studies introduces inconsistencies and complicates the interpretation of meta-analytic results. Only Kulik and Kulik (1988), which included a small subset of studies using computer-based feedback, attempted a more precise differentiation between delays by categorizing them as either post-item or post-test. Their findings indicated that when delayed feedback was given only a few seconds before the next item, immediate feedback was significantly more effective. However, when feedback

was postponed until after an entire test, this advantage disappeared, with delayed feedback even leading to slightly better performance. Despite this attempt at differentiation, further refinement is needed to capture the spectrum of feedback timings used in digital learning environments.

Secondly, there is a noticeable lack of recent meta-analyses that directly compare immediate and delayed feedback within the context of modern digital learning environments. Earlier reviews like those by Azevedo and Bernard (1995), Van der Kleij et al. (2015), and Swart et al. (2019) primarily used feedback timing as a moderating factor in meta-analyses of feedback effects, which may obscure specific effects attributable to timing. The most recent direct comparison of immediate and delayed feedback goes back to Kulik and Kulik (1988) and therefore misses nearly four decades of studies on feedback timing.

The Present Study

Given the inconsistent findings regarding the optimal timing of feedback, limitations in previous meta-analyses, rapid technological advancements in education, and the significant expansion of empirical research on feedback timing, there is a pressing need to re-examine and synthesize the literature on feedback timing within digital learning settings. This meta-analysis aims to clarify the effectiveness of different feedback timings by focusing specifically on studies that directly compare immediate feedback with delayed feedback. The key research questions are:

- **RQ1:** To what extent does feedback timing (immediate vs. delayed) influence learning outcomes?
- **RQ2:** How do varying definitions of "immediate" and "delayed" feedback influence the reported effectiveness of these feedback timing effects on learning outcomes?
- **RQ3:** What other factors moderate the effects of immediate versus delayed feedback on learning outcomes?

Methods

Our meta-analysis was conducted in accordance with the PRISMA guidelines (Page et al., 2021) to address the defined research questions. The methodology included pre-registered[1] inclusion and exclusion criteria, a structured search protocol, a systematic screening protocol, data extraction protocols, and detailed statistical methods.

Inclusion and Exclusion Criteria

In line with previous review studies on the effects of feedback timing on learning, and with our research objective to calculate the aggregate effect of immediate versus delayed feedback on learning, we defined the following inclusion criteria:

[1] Pre-registration available at OSF repository. Data and materials available at OSF Files page.

(a) **Publication Type:** Studies must be published in peer-reviewed scientific journals, unpublished doctoral dissertations, or in conference proceedings.
(b) **Publication Year:** Studies must be published after 1988. This cutoff was set to include the corpus of studies not reviewed in the last major meta-analysis (Kulik & Kulik, 1988).
(c) **Publication Language:** Publications must be in English, Turkish, or French, covering a broad spectrum of educational research within our linguistic capabilities.
(d) **Educational Context:** Studies must involve meaningful learning processes, such as acquiring new knowledge, skills, or concepts, regardless of whether they occur in formal education settings or controlled experimental environments designed to promote learning.
(e) **Learning Environment:** Studies should be conducted in computer-assisted settings to standardize feedback delivery and constrain the delay definition. This includes:

- Studies explicitly conducted within a computer-assisted environment.
- Studies utilizing digital communication platforms (e.g., Skype, Zoom) for experimental procedures.
- Experiments conducted in labs using digital tools or software.
- Any setting where the training and feedback components are delivered through computer systems, irrespective of the format of assessments (digital or directly by an instructor).

(f) **Participant Type:** Participants must be typical learners without special educational needs to ensure consistency in learning capabilities across studies. Studies are assumed to qualify whenever no specific disorders or disabilities are mentioned.
(g) **Feedback Definition:** The feedback operationalized in the study should comply with the following criteria:

- Feedback must be provided on an individual basis, directly related to participants' responses or performances.
- Feedback timing should be clearly defined as either immediate or delayed. Studies using alternative terminology (e.g., synchronous/asynchronous) that logically imply a comparison between immediate and delayed feedback also meet this criterion.
- Although exact times or item counts are not mandatory, the description should categorize the feedback as item-based or time-based, and providing an indication of the implemented delays.

(h) **Independent Variable:** Studies must directly compare the effects of immediate versus delayed feedback, addressing our central research question.
(i) **Dependent Variable:** The study should focus on learning outcomes as a primary measure of success. Learning outcomes must be measurable, such as post-test scores or skill acquisition levels, to ensure an objective assessment of feedback effectiveness.

(j) **Study Design:** Studies must employ a randomized controlled experimental design to minimize potential biases and eliminate confounding variables. This includes studies where participants are randomly or pseudorandomly assigned to conditions that are identical except for the timing of the feedback, with no obvious confounding factors.
(k) **Statistics:** Studies must provide sufficient statistical data (e.g., means, standard deviations, t-values, F-values) to allow for the extraction or computation of effect sizes.

Deviations from Pre-registered Inclusion Criteria The above inclusion and exclusion criteria are those that were pre-registered, with four modifications: (1) We added a criterion related to the educational context to exclude studies focused on non-semantic learning, tasks without meaningful content or educational value, such as perceptual discrimination, as these were less relevant to our main purpose. (2) We refined the feedback definition criterion after encountering studies where feedback was based on rankings or behaviors not directly linked to task responses, or where feedback was summarized (e.g., number of correct responses) rather than tied to specific answers. (3) We refined the definition of a computerized environment to include specifically studies where training and feedback are delivered via computer systems, regardless of whether pretests and post-tests were computerized, thus focusing on the importance of the training phase. (4) We initially required detailed reporting of delayed feedback timing for the purpose of moderator analyses, but we relaxed it as it was not critical for calculating the general pooled effect size. This allowed the inclusion of two additional studies (Dunbar et al., 2017; Lavolette et al., 2015) which did not provide specific delay details but met all other inclusion criteria.

Search Protocol

To identify studies for inclusion in our meta-analysis, we conducted systematic searches across three major databases: Web of Science, Scopus, and ERIC. The comprehensive search terms are detailed in the supplementary appendix (see Appendix A), and the dataset of retrieved papers is available on the OSF Files page (see footnote 1).

To ensure the efficacy of our search strategy, we tested it against a core set of seminal papers (Attali & van der Kleij, 2017; Metcalfe et al., 2009; Nakata, 2015; Van der Kleij et al., 2012), which are pivotal in this research area. This validation step helped to confirm the inclusivity and accuracy of our search terms, allowing for necessary adjustments. Our database search yielded 6,422 papers. After removing 1,322 duplicates, we were left with 5,100 papers.

Additionally, we set up alerts in both Web of Science and Scopus to notify us of new publications relevant to our search terms, until December 6, 2024, when we froze the database. This feature ensured that we were informed weekly of the latest studies, allowing for timely inclusion in our analysis. We have also examined the reference lists of eligible studies and relevant review papers identified during the initial search. This reverse search process and the inspection of the "relevant studies" section in Web of Science led to the inclusion of 8 additional records.

For an overview of the search steps and the corresponding numbers of articles at each stage, refer to the PRISMA diagram in Fig. 1.

Screening Protocol

The selection of included studies involved a methodical two-stage screening process, following initial retrieval from our database searches. These stages are summarized in Fig. 1.

The initial screening was based solely on information available in the title and abstract, focusing on criteria such as study focus and methodology. The first author conducted this screening, which led to the exclusion of 4,945 papers. To verify the accuracy and consistency of this screening, approximately 10% of these records (n = 500) were independently reviewed by a second screener. Inter-rater reliability was substantial, with a Kappa coefficient of 0.742 (95% CI [0.725, 0.759]), indicating a high level of agreement (Landis & Koch, 1977). Cases of disagreement (n = 8) were discussed and resolved jointly. In 7 out of these 8 cases, the final decision matched the initial judgment of the first screener.

The remaining 155 papers underwent a full-text review against all inclusion and exclusion criteria, along with a quality assessment conducted using the tool proposed by Kmet (2004). This tool, which evaluates 14 different criteria, is designed to assess the rigor of quantitative research through the adequacy of research questions, study design appropriateness, and thoroughness in result reporting. Three studies (Schroth, 1992, 1995; Schroth & Lund, 1993), contributing a total of 21 effect sizes, were excluded during this quality assessment stage due to critical methodological confounds. Although they met the core inclusion criteria, their training procedures required participants to reach a performance threshold (e.g., 9 out of 10 correct) rather than undergoing a fixed number of trials. As reported by the authors, this

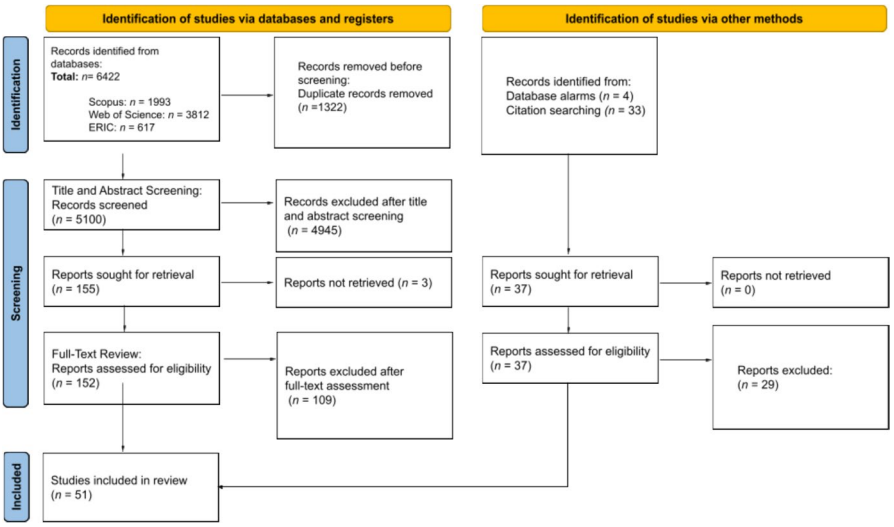


Fig. 1 PRISMA flow diagram based on the PRISMA 2020 model

resulted in substantially more training for participants in delayed feedback condition. Thus, while they concluded that delayed feedback enhanced transfer, this advantage is confounded by unequal training exposure, making it unclear whether the effect is attributable to feedback timing or additional practice.

After excluding a total of 109 papers in this full-text review stage and an additional 3 that could not be retrieved, 43 papers were found eligible for data extraction. Approximately 13% of the 155 papers ($n = 20$) were independently reviewed to assess reliability, achieving a Kappa value of 0.70 (95% CI [0.510, 0.890]), indicating a high degree of consistency (Landis & Koch, 1977). A detailed record of reasons for exclusion at each screening stage is available on the OSF Files page (see footnote 1).

With the addition of 8 papers found through other methods, we ended up with 51 studies from which the effect size was extracted and moderator variables were coded.

Data Extraction

Extraction of Statistics for Calculating Effect Sizes

The procedures for calculating effect sizes is detailed in the Statistical Methods section. To ensure accurate computation, we followed several guidelines during the initial coding of relevant statistics:

(a) To ensure consistency, the standardized mean differences were always coded with delayed feedback as the reference group. Thus, a positive mean difference indicates that immediate feedback is more effective than delayed feedback, and a negative effect size suggests the opposite.
(b) Following the recommendations by Lipsey (2001), priority was given to directly available statistics such as means and standard deviations. When these were not available, we used other statistics (e.g., t-tests, regression coefficients, ANOVA results) and transformed them according to established formulas.
(c) When studies reported statistics for multiple conditions, we generally coded them separately unless the conditions were not relevant to our research questions or moderators.
(d) If studies reported both raw and adjusted learning outcomes (adjusted for prior knowledge, usually using pre-test results), the adjusted outcomes were coded. Otherwise, statistics were based on raw data. This was also recorded as a moderator in order to test whether this made a difference.

The initial coding of these statistics was conducted by the first author. To validate the reliability of this coding, two additional coders independently coded 10% of the included studies each, totaling 20% of the studies being double-coded. These coders were also tasked with calculating effect sizes based on their extracted statistics. Discrepancies between primary and secondary coders were resolved through discussion, with no significant issues affecting the final dataset. The correlation of effect sizes between the double-coded studies was $r = 0.96$ for 32 effect sizes.

Coding of Study Characteristics

In order to address inconsistencies in previous findings and conduct moderator analyses, we coded additional characteristics for each included study. As a starting note, "immediate" and "delayed" feedback were coded according to how each study originally defined them, rather than using a uniform threshold. This preserved the intended contrast within each study and informed our delay difference moderator.

The pre-registered moderators were the following:

Delay Unit Feedback delay unit was categorized into two types: item-based and time-based. In an item-based delay, the interval between an answer and its feedback is determined by the number of intervening items (e.g., 3 items). In a time-based delay, the interval is determined by the amount of time that elapses between an answer and the corresponding feedback (e.g., 10 seconds).

Delay Difference The delay difference between immediate and delayed feedback, in seconds or in the number of items, depending on the delay unit.

Feedback Type Drawing on the categorization proposed by Shute (2008), feedback was classified into four primary types: Verification, Correct Response, Try Again, and Elaborated. For simplicity, Verification and Correct Response were combined into a single category termed "Simple Feedback Type."

Education Level Categorized into Primary, Secondary, Tertiary, and Adult Education levels.

Learning Domain Studies were classified into specific learning domains:

- Text Memory: Tasks involving memorization of text for later recall.
- Language Learning: Tasks related to language skills, including vocabulary acquisition and memorization.
- Reading Comprehension: Focused on understanding and analyzing text.
- STEM: Focused on science, technology, engineering, and mathematics.
- Social Sciences: Includes studies on social behaviors, cognition, and societal issues.
- General Cognitive Skills: Encompasses cognitive tasks outside of language or STEM categories, such as probabilistic learning and general memory tasks.
- Mixed: Studies spanning multiple domains or interdisciplinary approaches.

Training Task Complexity Using Bloom et al. (1964) taxonomy, tasks were classified as Lower-Order (involving knowledge, comprehension, or recall) and Higher-Order (requiring application, analysis, synthesis, evaluation, or extended thinking). Studies encompassing both were categorized as Higher-Order.

Prior Knowledge Adjustment Indicates whether the effect sizes were adjusted for participants' prior knowledge and skills, as assessed through pretests administered before training.

Learning Task Context Experimental Tasks: Designed specifically for research, often detached from practical educational applications. Curriculum-Based Learning Tasks: Integrated within an actual curriculum, enhancing learners' educational or professional objectives.

In addition to the pre-registered moderators, several other study characteristics were coded. The impact of these additional characteristics was also explored in our moderator analysis. Below is a list of these characteristics along with their respective coding schemas:

- Participant Design: within-subject / between-subjects.
- Learning Task Type: classified as arbitrary / non-arbitrary, based on the meaningfulness of the learned associations. Arbitrary tasks involve learning associations that do not have inherent semantic connections (e.g., pairing unrelated words or symbols), while non-arbitrary tasks involve meaningful or conceptually related content (e.g., reading comprehension, solving math problems).
- Post-test Item Similarity: identical to / different from training items.
- Feedback Dependency on Errors: feedback provided only for incorrect responses / systematically.
- Feedback Answer Reminder: Indicates whether the feedback included a reminder of the initial answer or not.
- Number of Training Items: The total count of items used during training sessions.
- Post-test Task: memory retrieval/ knowledge application
- Time Limitation for Responses: time-limited / no time limit for answering during tasks.
- Retention Interval Difference (between Last Feedback and Post-test): The time difference, in seconds, between the last feedback and the post-test between immediate and delayed feedback conditions. Calculated as the retention interval of delayed feedback subtracted from that of immediate feedback.
- Publication Year: The year the study was published.

Deviations from Preregistered Coding Procedure Our analysis deviated from the preregistered coding procedure by combining the 'Verification' and 'Correct Response' feedback types into a new category named 'Simple Feedback Type.' This adjustment, not specified in the preregistration, was made to simplify the analysis and clarify impact distinctions between simpler and more complex feedback types. Additionally, this grouping was implemented to increase the number of effect sizes per group, enhancing the statistical robustness of the moderator analysis.

Statistical Methods

Calculation of Effect Sizes

In this meta-analysis, each effect size indicates the standardized difference in learning outcomes between groups under different feedback timing conditions, typically assessed using post-test scores.

To calculate the effect size we used the the data coded in the data extraction section and followed these formulas:

Cohen's d was calculated by:

$$d = \frac{M1 - M2}{S}$$

where $M1$ and $M2$ are the means of the immediate feedback and delayed feedback groups, respectively, and S denotes the pooled standard deviation. The pooled standard deviation was determined differently based on the study design:

For within-subject designs:

$$S = \sqrt{\frac{s1^2 + s2^2}{2}}$$

where $s1$ and $s2$ are the standard deviations for each of the two feedback timing conditions.

For between-subject designs:

$$S = \sqrt{\frac{(n1-1)s1^2 + (n2-1)s2^2}{n1 + n2 - 2}}$$

where $s1$ and $s2$ are the standard deviations for each of the two feedback timing groups and $n1$ and $n2$ are the numbers of participants in each group.

The standard error of the effect size ($d.se$) was again calculated based on the study design.

For within-subject designs:

$$d.se = \sqrt{\frac{2(1-r)}{n} + \frac{d^2}{2n}}$$

For between-subject designs:

$$d.se = \sqrt{\frac{n1 + n2}{n1 \times n2} + \frac{d^2}{2(n1 + n2)}}$$

Here, r denotes the within-subject correlation between the conditions and was assumed to be 0.5 if not reported. As is standard in meta-analyses, standard errors

were based on d, which was then adjusted to Hedges' g to correct for small sample bias:

$$g = d \times \left(1 - \frac{3}{4N - 9}\right)$$

where N is the total sample size encompassing all conditions. All reported effect sizes use Hedges' g.

Consistent with our coding rule, positive values of Hedges' g indicate higher learning outcomes for immediate feedback relative to delayed feedback, whereas negative values indicate an advantage for delayed feedback.

In cases where the necessary statistics were not directly accessible, alternative statistical measures such as t values, univariate F values, or χ^2 statistics along with sample sizes were employed. These were converted into effect sizes using established conversion formulas from the meta-analysis methodology literature (see Lipsey 2001 for detailed formulas).

Outlier Detection and Publication Bias

We assessed potential outliers in the final effect sizes (Hedges' g) by identifying any deviations exceeding three standard deviations from the mean effect sizes of several groups. This method aims to pinpoint effect sizes that might disproportionately influence the summary statistics or introduce bias. This approach was adapted from Van der Kleij et al. (2015), who explored feedback effects in computer-based learning environments.

To evaluate publication bias in our meta-analysis, we conducted Egger's regression test (Egger et al., 1997), which formally assesses data asymmetry and offers a robust indication of potential biases within the included studies. In addition, we included a funnel plot for visual inspection and transparency.

Computation of Weighted Mean Effect Sizes

All analyses reported in this study used Hedges' g for effect size measurement. We used the Robust Variance Estimation (RVE) method to address the common issue of multiple and non-independent effect size estimates arising from the same study (Hedges et al., 2010). This approach is preferred over traditional random effects models, which are not suitable for meta-analyses involving dependent effect size estimates. The RVE method is particularly advantageous as it provides a more accurate estimation of standard errors, leading to narrower confidence intervals for the weighted mean effect sizes (Hedges et al., 2010).

We applied the RVE model using the R software package *robumeta* (Fisher & Tipton, 2015). This package requires specifying the correlation (ρ) between within-study effects. We used the default setting of $\rho = 0.80$ for consistency. In order to evaluate the potential impact of ρ on both the mean effect size and the estimated between-study heterogeneity (I^2), we conducted sensitivity analyses. These analyses varied the correlation of within-study effect sizes from $\rho = 0.0$ to $\rho = 1.0$, allowing us to

explore the robustness of our findings across different assumptions (Tanner-Smith & Tipton, 2014).

Moderator Analysis

In accordance with our pre-registered analysis plan, we conducted separate meta-regressions for each moderator. These regressions were integrated into the multilevel meta-analytical models by including each moderator as a fixed effect.

Multiple Meta-regression

To complement the individual analyses of moderators, we conducted a multiple meta-regression to assess the influence of multiple moderator variables while controlling for the effects of others (Rubin, 1992). We only included the moderators that demonstrated significant effects in the separate analyses. To enhance statistical power, some moderator categories were consolidated by grouping their categories.

Deviations from Preregistration The multiple meta-regression was not part of our preregistration but was introduced because the significant moderators appeared to be confounded with each other, raising the question of which of those actually influenced the meta-analytic effects.

The complete set of analysis codes, along with the data utilized in this study, is accessible on the OSF Files page (see footnote 1)

Results

Descriptive Statistics

This meta-analysis includes 51 studies published from 1995 to 2024, collectively contributing a total of 160 effect sizes. A full list of the 51 included studies, along with descriptive information and coded preregistered moderator variables, is presented in Table B1 (see Appendix B). A more detailed data extraction, including exploratory coding of additional study characteristics, is available in the preregistered file on the OSF Files page.

On average, each paper reported approximately 3.14 effect sizes, with a maximum of 12 and a median of 2 effect sizes per study. Two papers, accounting for six effect sizes, originated from dissertations, while the remaining 49 papers, contributing 154 effect sizes, were sourced from peer-reviewed journal articles. The geographical distribution of studies and effect sizes is summarized in Table 1.

The included studies had a cumulative sample size of 14,313 participants. The average sample size per paper was approximately 117 ($SD = 338.39$), ranging from a minimum of 17 to a maximum of 2,445 participants. The average percentage of female participants was 58.96% ($SD = 20.04$), with a range from 0% to 96%. Gender distribution data were unavailable for 23 papers. The average reported age of

Table 1 Geographical distribution of studies and effect sizes

Country	Num. studies *N*	Num. effects *k*
United States	28	87
Japan	5	20
Spain	3	4
Netherlands	3	7
Germany	3	7
China	3	13
Australia	1	8
Taiwan	1	2
Saudi Arabia	1	3
Greece	1	2
Unspecified	2	4

participants was 21.39 years ($SD = 10.51$), with ages ranging from 8.2 to 64.4 years. Age data were not reported in 27 papers. The majority of the studies (76.5%) focused on university students.

Among studies that defined delay in seconds (time-based), the mean delay for immediate feedback was $M = 0.03$ seconds ($SD = 0.12$ seconds, median = 0.00 seconds, range = 0.00–0.50 seconds). This low mean reflects that most immediate feedback conditions were coded as 0 seconds, indicating feedback delivered immediately upon response. For delayed feedback, the geometric mean (because times were log-transformed) was approximately 559.73 seconds ($\approx$ 9.33 minutes; $SD = 165{,}545.87$ seconds, $\approx$ 46 hours; median = 525.00 seconds, $\approx$ 8.75 minutes; range = 1.00–604,800.00 seconds $\approx$ 7 days).

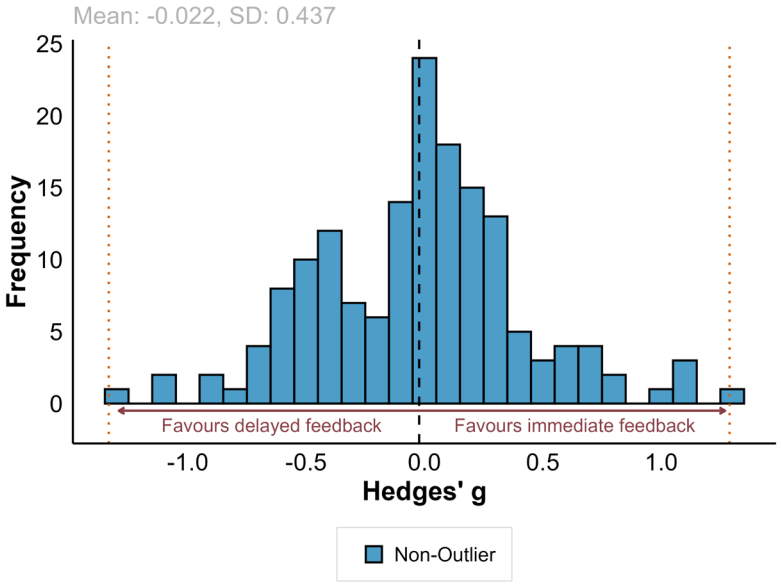
Among studies that defined delay in terms of item count, the mean delay for immediate feedback was $M = 1.18$ items ($SD = 0.87$, median = 1.00, range = 0.00–5.00), and for delayed feedback it was $M = 20.74$ items ($SD = 13.18$, median = 14.50, range = 3.00–60.00).

To complement the descriptive statistics, Table C2 (see Appendix C) summarizes the included studies by delay definition and categorizes delayed feedback durations as short, intermediate, or long, reporting the number of studies and effect sizes in each group. Overall, the included studies encompassed a wide range of feedback delays. As shown in Table C2, delays in delayed feedback condition ranged from 1 second to 7 days in time-based designs and from 1 to 60 intervening items in item-based designs. Approximately 45% of effects involved short delays (less than one hour or up to 10 items), 30% involved intermediate delays (one hour to one day or 11–30 items), and 25% involved long delays (more than one day or more than 30 items). Thus, although brief delays were common, the dataset also included a substantial subset of studies employing multi-hour or $\geq$ 10-item feedback intervals.

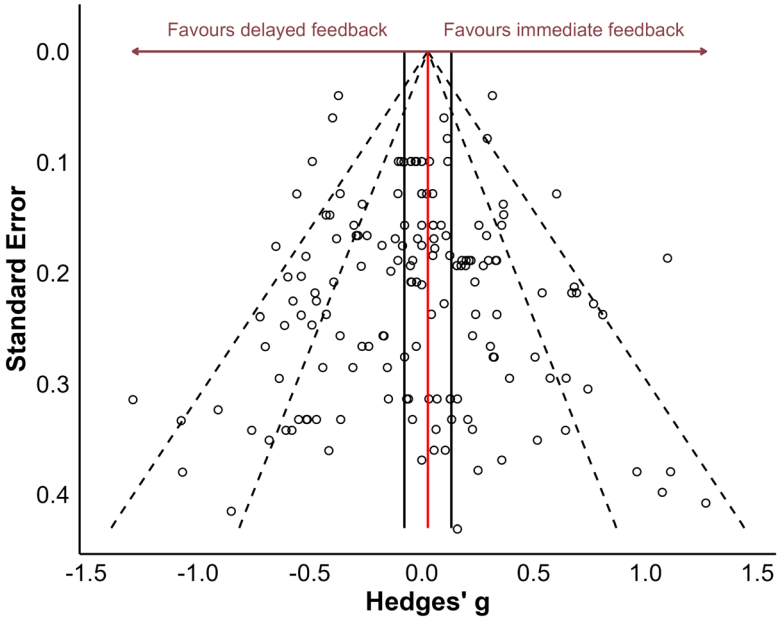
Detailed study characteristics that can serve as potential moderators are presented in the "Moderators" section.

Outlier Detection and Publication Bias

The results of the outlier diagnostics, as illustrated in Fig. 2a, indicate that no effect sizes deviated by more than three standard deviations from the mean group effect size of -0.022. Consequently, no effect sizes were removed at this stage.



(a) Outlier Check Histogram



(b) Funnel Plot

Fig. 2 (**a**) Outliers check histogram and (**b**) Funnel plot

To evaluate potential publication bias, we first visualized the distribution of effect sizes against standard errors via a funnel plot, which appeared largely symmetrical (Fig. 2b). Egger's regression test yielded an intercept of 0.09 ($SE = 0.32$, $t = 0.28$, $p = .78$), indicating no significant deviation of the expected effect sizes from zero when precision is low. Similarly, the regression coefficient for precision (inverse SE) was -0.04 ($SE = 0.05$, $t = -0.79$, $p = .43$), suggesting that effect sizes were not systematically related to study precision. Together, these results indicate no significant evidence of publication bias in this meta-analysis, and therefore no adjustments for publication bias were necessary.

Overall Effect of Immediate versus Delayed Feedback on Learning (RQ1)

Using the Robust Variance Estimation (RVE) method, we calculated the overall weighted mean effect size from the 160 individual effect size estimates (from 51 studies), yielding $g = 0.03$ (95% CI [-0.08, 0.13], $p = .61$) with an estimated between-study standard error of 0.05. Heterogeneity was substantial ($I^2 = 82\%$), and the between-study variance was notably high ($\tau^2 = 0.12$), suggesting that study conditions or populations may influence the effect of feedback timing and motivating subsequent moderator analyses.

Sensitivity Analyses

Modifying the assumed within-study effect size correlation (ρ) did not affect the estimated value of g (see Appendix Table D3). The leave-one-study-out influence analysis showed that the overall effect estimate ranged from $g = 0.00$ to $g = 0.04$ when individual studies were sequentially removed, indicating robust findings not driven by any single study (see Appendix Fig. E1 for detailed results).

To evaluate whether very brief delays biased the overall estimate, we conducted a sensitivity analysis excluding all effects with delays shorter than one minute (for time-based designs) or fewer than five items (for item-based designs). Based on 133 effect sizes from 43 studies, the overall effect remained non-significant ($g = 0.00$, 95% CI [-0.10, 0.10], $p = .95$), indicating that the inclusion of very short delays did not materially influence the results.

To further examine whether delay length moderates the timing effect, we conducted a categorical meta-analysis using the predefined short, intermediate, and long delay groupings from Table C2 (< 1 hour or 1–10 items; 1 hour–1 day or 11–30 items; > 1 day or > 30 items). The pooled estimates and heterogeneity values for each subgroup are presented in Table 2 and visualized in the subgrouped forest plot (Fig. 3). Across categories, the effects remained small and non-significant, including the long-delay subset ($k = 31$, $g = -0.06$, 95% CI [-0.26, 0.15], $p = .52$), indicating that even extended feedback intervals did not produce a reliable advantage over immediate feedback.

Finally, to evaluate the robustness of the findings under delay intervals most relevant to classroom practice, we conducted a targeted subset analysis restricted to studies implementing feedback delays of at least one day. This subset comprised 7 studies contributing 17 effect sizes. The pooled effect was moderate in magnitude

Table 2 Meta-analytic results by delay category

Delay category	Est.	p	95% CI	τ^2	I^2	Studies N	Effects k
Short	0.13	.15	[-0.05, 0.32]	0.12	82%	21	66
Intermediate	-0.05	.60	[-0.26, 0.15]	0.22	88%	19	45
Long	-0.06	.52	[-0.26, 0.15]	0.06	66%	9	31
Long ($\geq$1 day)	-0.35	.09	[-0.78, 0.07]	0.26	85%	7	17
Overall	**0.03**	**.61**	**[-0.08, 0.13]**	**0.12**	**82%**	**51**	**160**

Signif. codes: < .01 ** < .05 *

Note. Short: <1 hour or 1–10 items; Intermediate: 1 hour–1 day or 11–30 items; Long: >1 day or >30 items. Long (>1 day) and Long ($\geq$1 day) represent targeted subset analyses restricted to classroom-relevant delays. $\rho = 0.80$

but did not reach statistical significance ($g = -0.35$, 95% CI [-0.78, 0.07], $p = .09$), with substantial heterogeneity across studies ($I^2 = 85\%$, $\tau^2 = 0.26$). These results are presented in Table 2 (Long ($\geq$1 day) category), and the corresponding forest plot for this subset analysis is shown in Appendix Fig. F2.

Moderators

Moderating Effect of Definitions of "Immediate" and "Delayed" Feedback (RQ2)

To preserve the intended contrast within each study, "immediate" and "delayed" feedback were coded based on the definitions utilized within each original study, and the differences between these definitions was used to construct the delay difference moderator. Our reading of the literature has identified two ways in which the definition of feedback delay varied: the first was the unit in which the delay was defined (absolute time, number of intervening items, or a combination of both), and the second was the delay difference (between delayed and immediate feedback, which is itself sometimes delivered after a non-null delay).

Comparing delays defined in seconds (num. effects $k = 66$) to those defined in items ($k = 92$, reference category) revealed no significant difference ($g = -0.12$, $p = .20$). Studies using either definition showed no significant advantage for a specific feedback timing, with effect sizes of $g = -0.09$ ($p = .24$) for time-based definitions and $g = 0.03$ ($p = .57$) for item-based definitions. A third category combining time- and item-based delay definitions is also shown in Table 3, but it is not interpreted due to the limited number of effect sizes ($k = 2$).

Analyses of the moderating effect of delay difference were first carried out separately for studies defining feedback delay by time and those defining it by items. Since delay differences measured in time covered several orders of magnitude (from seconds to days) and tended to be highly skewed, we applied a log transformation to normalize their distribution and reduce the influence of extreme values. For studies defining delay by time ($k = 62$ where the delay difference was reported), there was no significant effect of delay difference ($g = -0.02, p = .35$), (see section Delay Difference (Seconds(log)) in Table 3).

For the studies defining delay in items ($k = 80$ where the delay difference was reported), there was a significant effect of delay difference ($g = -0.01, p = .03$), (see

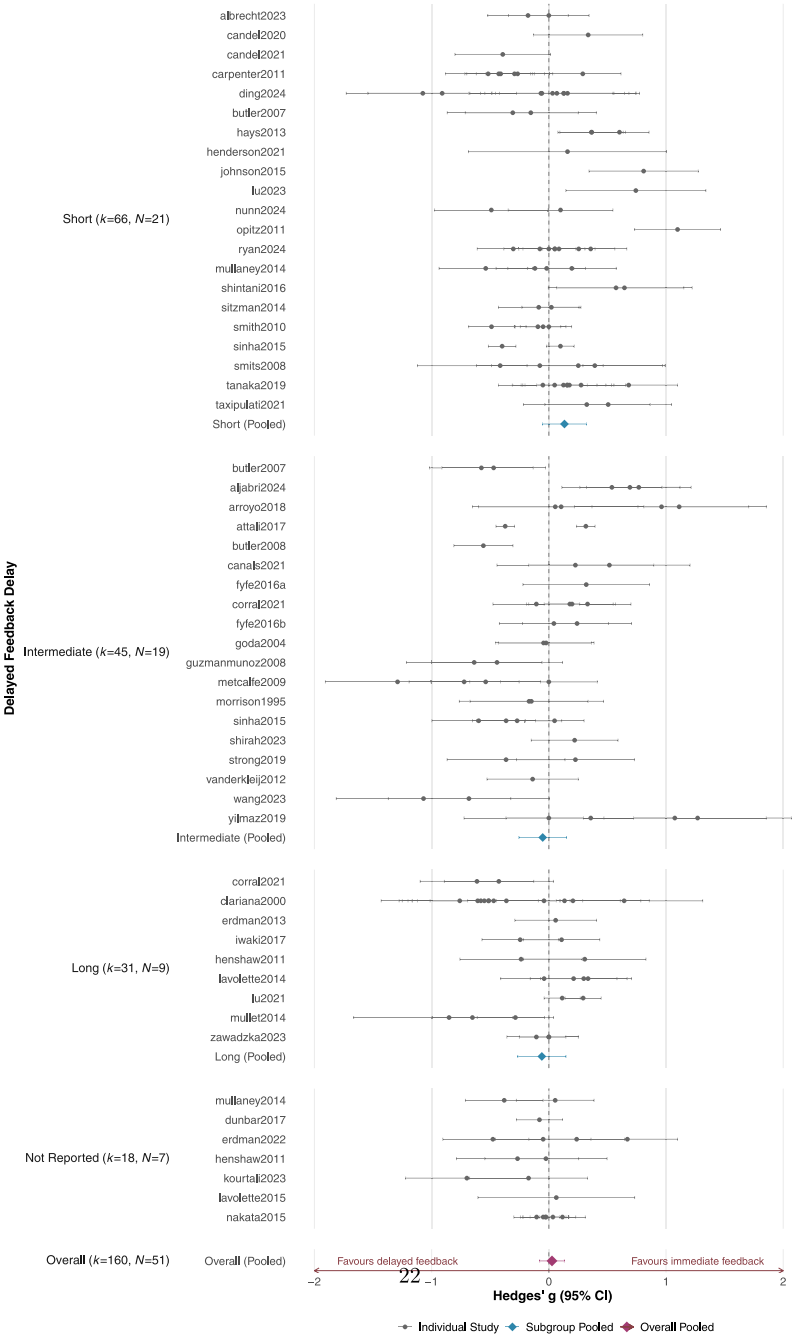


Fig. 3 Forest plot of effect sizes (Hedges' g) by feedback delay category. Individual effects (gray circles), subgroup pooled estimates (blue diamonds), and overall pooled effect (purple diamond). Short delays: < 1 hour or 1–10 items; Intermediate delays: 1 hour–1 day or 11–30 items; Long delays: > 1 day or > 30 items

Table 3 Preregistered moderator effects on feedback timing

Moderator	Estimate	Std. error	*t*	*p*	95% CI	Num. studies *N*	Num. effects *k*
Delay unit						51	160
Intercept (Item)	0.03	0.06	0.57	.57	[-0.08, 0.14]		92
Seconds (vs. item)	-0.12	0.09	-1.29	.20	[-0.31, 0.07]		66
Mixed (item+seconds) (vs. item)	-0.18	0.06	-3.21	.002**	[-0.29, -0.07]		2
Delay difference (Seconds, log)						18	62
Intercept (delay difference = 1 s.)	0.03	0.14	0.19	.85	[-0.28, 0.33]		
Delay difference	-0.02	0.02	-0.96	.35	[-0.06, 0.02]		
Delay difference (Items)						29	80
Intercept (Delay difference = 2 items)	0.19	0.09	2.09	.05*	[0.00, 0.38]		
Delay difference	-0.01	0.00	-2.36	.03*	[-0.02, 0.00]		
Delay difference (Items converted to seconds, log)						16	54
Intercept (delay difference = 1 s.)	0.26	0.45	0.57	.58	[-0.72, 1.23]		
Delay difference	-0.03	0.06	-0.53	.60	[-0.17, 0.10]		
Delay difference (All converted to seconds, log)						34	116
Intercept (delay difference = 1 s.)	0.06	0.14	0.43	.67	[-0.23, 0.36]		
Delay difference	-0.02	0.02	-0.79	.44	[-0.06, 0.03]		
Feedback type						51	160
Intercept (Elaborated)	-0.03	0.06	-0.50	.62	[-0.15, 0.09]		40
Simple (vs. elaborated)	0.02	0.08	0.20	.84	[-0.15, 0.18]		118
Try-again (vs. elaborated)	-0.37	0.06	-6.17	< .001**	[-0.49, -0.25]		2
Educational level						51	160
Intercept (Tertiary)	0.01	0.05	0.30	.76	[-0.08, 0.11]		127
Secondary (vs. tertiary)	-0.25	0.11	-2.15	.04*	[-0.47, -0.02]		22
Adult education (vs. tertiary)	-0.10	0.06	-1.73	.09	[-0.23, 0.02]		6
Primary (vs. tertiary)	-0.23	0.32	-0.70	.49	[-0.87, 0.42]		5
Educational level (Regrouped)						51	160
Intercept (Tertiary + Adult)	0.01	0.04	0.19	.85	[-0.08, 0.10]		133
Primary + Secondary (vs. Tertiary + Adult)	-0.23	0.11	-2.09	.04*	[-0.46, -0.01]		27

Table 3 (continued)

Moderator	Estimate	Std. error	*t*	*p*	95% CI	Num. studies *N*	Num. effects *k*
Learning domain						51	160
Intercept (Text memory)	-0.30	0.06	-5.02	< .001**	[-0.41, -0.18]		18
STEM (vs. text memory)	0.24	0.09	2.71	.01**	[0.06, 0.43]		35
Social science (vs. text memory)	0.14	0.06	2.43	.02*	[0.02, 0.26]		3
Reading comprehension (vs. text memory)	-0.57	0.06	-9.71	< .001**	[-0.69, -0.45]		2
Language learning (vs. text memory)	0.42	0.10	4.02	< .001**	[0.21, 0.63]		53
General cognitive skills (vs. text memory)	0.25	0.09	2.66	.01*	[0.06, 0.44]		49
Learning domain (Regrouped)						51	160
Intercept (Other domains)	0.01	0.04	0.27	.79	[-0.08, 0.10]		20
Text-based learning (vs. other)	-0.35	0.08	-4.19	< .001**	[-0.52, -0.18]		140
Training task Complexity						51	160
Intercept (Higher order)	0.04	0.06	0.58	.56	[-0.09, 0.16]		76
Lower order (vs. higher order)	-0.10	0.09	-1.19	.24	[-0.28, 0.07]		84
Prior knowledge Adjustment						51	160
Intercept (Adjusted)	0.10	0.14	0.73	.47	[-0.17, 0.37]		25
Not adjusted (vs. adjusted)	-0.14	0.14	-0.96	.34	[-0.43, 0.15]		135
Learning task context						51	160
Intercept (Curriculum-based)	0.04	0.06	0.56	.58	[-0.09, 0.16]		58
Experimental (vs. curriculum-based)	-0.09	0.08	-1.08	.29	[-0.26, 0.08]		102

Signif. codes: < .01 ** < .05 *

Note. For moderators significant in the initial analysis and with more than two levels, grouping was applied to simplify the multiple meta-regression (see Section "Other Preregistered Moderators (RQ3)"). 'Text-based learning' combines text memory and reading comprehension, which both showed negative coefficients relative to the reference, indicating similar delayed-feedback patterns, whereas all other domains (STEM, language learning, social sciences, and general cognitive skills) showed positive coefficients, suggesting relatively greater immediate-feedback effects

section Delay Difference (Items) in Table 3). This means that the effect of delayed feedback on performance, although negative for small delay differences ($g = 0.19$ for 2 items of delay, meaning better performance for immediate feedback), increased by 0.01 standard deviations by item of delay. Thus, when the delay difference between immediate and delayed feedback increased beyond 21 items, it started having positive effects on performance (see Fig. G3c).

In order to test the effect of delay difference across a maximum of studies irrespective of their definition of feedback delay, we attempted to convert item-based delays into equivalent times, for studies providing enough information on item timing to do so. This conversion was possible for 54 effects. For this subset of studies, where delays were converted from items to seconds and subsequently log-transformed, the moderating effect of delay difference was no longer significant ($g = -0.03$, $p = .60$), (see section Delay Difference (Items Converted to Seconds(log)) in Table 3).

Finally, when considering all studies expressing feedback delay in time or converted from items to seconds, totaling 116 effects (62 originally reported in seconds and 54 converted from items), there was no significant moderating effect of delay difference on learning outcomes ($g = -0.02$, $p = .44$), (see section Delay Difference (All Converted to Seconds(log)) in Table 3).

In conclusion, we found evidence in favor of the effect of delayed feedback on performance only in studies defining delay in items, and using a large number of intervening items.

Other Preregistered Moderators (RQ3)

In addressing our third research question, we first evaluated the impact of preregistered moderators on the comparative effectiveness of immediate versus delayed feedback on learning outcomes. The distributions of effect sizes for the eight preregistered moderators are comprehensively detailed in the supplementary appendix, as illustrated in Fig. G3.

Among the examined moderators, no significant difference was found when comparing simple feedback ($k = 118$) to elaborated feedback ($k = 40$, reference category), ($g = 0.02$, $p = .84$; see Table 3). This suggests that the type of feedback used in the experiment—whether simple or elaborated—does not significantly influence the effect of feedback timing on learning outcomes. An additional category of try-again feedback is included in Table 3, but it is not interpreted here due to the limited number of effect sizes ($k = 2$).

A significant moderating effect was observed when comparing studies conducted at the secondary ($k = 22$) with those at the tertiary education levels ($k = 127$, reference category), ($g = -0.25$, $p = .04$). This suggests that studies involving secondary education participants demonstrated greater improvements in performance under delayed feedback conditions ($g = -0.23$, $p = .03$), whereas studies with tertiary education participants showed no significant advantage for either immediate or delayed feedback ($g = 0.01$, $p = .76$). Additionally, again using studies at the tertiary education level as the reference category, studies conducted in adult education ($k = 6$) and primary education ($k = 5$) showed a nominal but non-significant advan-

tage for delayed feedback ($g = -0.10$, $p = .09$) and ($g = -0.23$, $p = .49$), respectively, based on relatively few numbers of effects.

For the purpose of the subsequent meta-regression, we grouped together similar education levels into just 2 categories: 1) primary + secondary and 2) tertiary + adult. This grouped version of the moderator was re-tested and revealed that delayed feedback significantly improved learning outcomes in studies involving participants in primary or secondary education compared to those in tertiary or adult education ($g = -0.23$, $p = .04$), (see section Educational Level (regrouped) in Table 3).

The moderator analysis revealed a differential impact of feedback timing across learning domains. Compared to experiments in which the domain was text memorization ($k = 18$, reference category), studies on general cognitive skills ($k = 49$, $g = 0.25$, $p = .01$), language learning ($k = 53$, $g = 0.42$, $p < .001$), and STEM ($k = 35$, $g = 0.24$, $p = .01$) demonstrated a significant advantage of immediate feedback.

Examining the absolute effect of feedback timing per learning domain, irrespective of the comparison, revealed that text memory ($g = -0.30$, $p < .001$) demonstrated a significant advantage for delayed feedback. Conversely, cognitive skills ($g = -0.05$, $p = .54$), language learning ($g = 0.13$, $p = .15$), and STEM ($g = -0.05$, $p = .47$) did not exhibit significant advantages in feedback timing. Other domains, such as reading comprehension and social sciences, are shown in Table 3 but are not interpreted here due to the limited number of effect sizes ($k = 2$–3).

For the subsequent meta-regression analysis, the learning domain moderator was restructured by combining reading comprehension and text memory into a single category labeled “text-based learning”, while all other domains were grouped under “others”. This revised moderator analysis revealed that, in experiments focusing on text-based learning, delayed feedback had a significant advantage compared to other domains ($g = -0.35, p < .001$), (see Section Learning Domain (grouped) in Table 3).

Regarding training task complexity, there was no significant difference in feedback delay effectiveness between lower-order tasks ($k = 84$) and higher-order tasks ($k = 76$), ($g = -0.10$, $p = .24$). Similarly, prior knowledge adjustment did not significantly moderate the effectiveness of feedback timing, as shown by the comparison of unadjusted ($k = 135$) to adjusted outcomes ($k = 25$), ($g = -0.14$, $p = .34$). Finally, learning task context did not moderate the impact of feedback timing, with no significant difference observed between experimental tasks ($k = 102$) and curriculum-based tasks ($k = 58$), ($g = -0.09$, $p = .29$).

Exploratory Moderators (RQ3)

Beyond the preregistered moderators, we examined ten exploratory moderators identified through the screening procedure. As shown in Table 4, only three were found to significantly moderate the effect of feedback timing on learning outcomes. Effect size distributions for all these exploratory moderators are provided in the appendix (Fig. H4)

A significant effect was observed when comparing studies without a time limit for responses ($k = 42$) to those with a limited response time ($k = 60$, reference category), ($g = -0.19$, $p = .03$). Studies that did not impose a time limit for responses

Table 4 Exploratory moderator effects on feedback timing

Moderator	Estimate	Std. error	*t*	*p*	95% CI	Num. studies *N*	Num. effects *k*
Participant design						51	160
Intercept (Between)	0.02	0.06	0.36	.72	[-0.10, 0.14]		94
Within (vs. between)	-0.09	0.08	-1.05	.30	[-0.26, 0.08]		66
Learning task type						51	160
Intercept (Arbitrary)	0.03	0.10	0.32	.75	[-0.17, 0.23]		30
Non-arbitrary (vs. arbitrary)	-0.07	0.11	-0.63	.53	[-0.29, 0.15]		130
Post-test item similarity						47	154
Intercept (Different from training)	0.04	0.07	0.58	.57	[-0.10, 0.18]		55
Identical to training (vs. different)	-0.10	0.09	-1.06	.30	[-0.28, 0.09]		80
Mixed (vs. different)	-0.07	0.18	-0.41	.69	[-0.43, 0.29]		19
Feedback dependency on errors						51	160
Intercept (Independent of error)	-0.03	0.04	-0.70	.49	[-0.12, 0.06]		126
Only for errors (vs. independent)	0.06	0.14	0.41	.69	[-0.23, 0.34]		34
Feedback answer reminder						35	109
Intercept (No answer reminder)	-0.04	0.08	-0.53	.60	[-0.22, 0.13]		65
Answer reminder (vs. no reminder)	0.02	0.11	0.17	.87	[-0.21, 0.25]		44
Number of training items						46	148
Intercept (Items = 4)	-0.03	0.06	-0.46	.65	[-0.14, 0.09]		
Number of items	0.00	0.00	0.00	1.00	[0.00, 0.00]		
Post-test task						51	160
Intercept (Knowledge application)	0.06	0.06	1.09	.28	[-0.05, 0.18]		74
Memory retrieval (vs. knowledge application)	-0.15	0.08	-1.82	.07	[-0.32, 0.02]		86
Time limitation for responses						51	160
Intercept (Limited)	0.04	0.07	0.64	.52	[-0.09, 0.18]		60
Not limited (vs. limited)	-0.19	0.08	-2.27	.03*	[-0.36, -0.02]		42
Not reported (vs. limited)	-0.03	0.11	-0.31	.75	[-0.25, 0.18]		58
Retention interval difference (log)						33	110
Intercept (Difference = 0)	0.05	0.08	0.70	.49	[-0.10, 0.21]		
Retention interval difference	-0.01	0.01	-1.39	.18	[-0.02, 0.00]		
Publication year						51	160
Intercept (Year = 1995)	-0.34	0.10	-3.40	.001**	[-0.55, -0.14]		
Publication year	0.02	0.01	3.22	.002**	[0.01, 0.03]		

Signif. codes: < .01 ** < .05 *

demonstrated improved performance under delayed feedback conditions ($g = -0.15$, $p = .005$), whereas studies with a limited response time showed no significant advantage for either immediate or delayed feedback ($g = 0.04$, $p = .52$).

Finally, the publication year of the studies emerged as a significant moderator ($g = 0.02$, $p = .002$), indicating that the effect of delayed feedback on performance, though positive in older studies (intercept at the minimum publication year, 1995, with $g = -0.34$, suggesting better performance under delayed feedback), decreased

by 0.02 standard deviations per year. It reached approximately zero around 2012, after which more recent studies found no benefit for delayed feedback (see Fig. H4j).

Multiple Meta-regression

In our multiple meta-regression, we included only the three statistically significant moderators: educational level, learning domain, and time limitation for responses. The publication year moderator, although significant in the individual analyses, was excluded, as it is not an experimental or feedback-related property but rather a factor likely to covary with other moderators.

To facilitate the interpretation of the multiple meta-regression, we used the grouped versions of certain moderators (educational level and learning domain) to ensure that only binary moderators were included.

No grouping was performed for the time limitation for response moderators. Notably, the "Not Reported" category in the time limitation for responses moderator was retained to account for the substantial number of effect sizes ($\mathrm{n} = 58$) with missing data. This inclusion ensured that the statistical power of the multiple meta-regression was not reduced by excluding these cases.

While individual analyses identified significant moderating effects for these three variables, the results of the multiple meta-regression (Table 5) indicate that none of the moderators remained statistically significant. This suggests that these moderators are partially confounded with one another.

To further explore the potential confounding relationships among the moderators included in the multiple meta-regression, we analyzed the distribution of educational level, learning domain, and time limitation for responses in relation to one another using a contingency table (Table 6). The table revealed notable dependencies between certain moderators. Specifically, educational level and learning domain were closely related ($\chi^2(1) = 41.76, p < .001$), as studies in tertiary/adult education primarily focused on non-text-based learning ($k = 127$), whereas those in primary/secondary education were more evenly distributed across learning domains.

Additionally, the publication year distribution of each moderator was examined (Fig. 4) to assess whether the publication year of the studies may have systematically influenced moderator effects. The distributions indicate a noticeable shift over time, with an increasing number of studies involving tertiary/adult participants ($t(30.81) = 4.65, p < .001, d = 1.28$), non-text-based learning tasks ($t(22.01) = -6.33, p < .001, d = -1.91$), and responses in a limited time ($t(73.89) = 4.24, p < .001, d = 0.89$).

Discussion

This meta-analysis systematically examined the impact of immediate versus delayed feedback on learning outcomes. It builds on previous research on feedback timing by systematically examining a broader range of potential moderating factors. One key advancement is its consideration of the variability in how studies define immediate and delayed feedback to address a significant gap in prior meta-analyses and

Table 5 Multiple meta-regression results

Moderator	Estimate	Std. error	*t*	*p*	95% CI
Intercept	-0.04	0.15	-0.24	.81	[-0.33, 0.26]
Educational level (Regrouped)					
Tertiary + Adult (vs. Primary + Secondary)	0.10	0.14	0.71	.48	[-0.18, 0.37]
Learning domain (Regrouped)					
Text-based learning (vs. others)	-0.26	0.14	-1.80	.08	[-0.54, 0.03]
Time limitation for responses					
Not limited (vs. limited)	-0.14	0.08	-1.73	.09	[-0.29, 0.02]
Not reported (vs. limited)	-0.03	0.10	-0.25	.81	[-0.23, 0.18]

Num. Studies $N = 51$; Num. Effects $k = 160$

Signif. codes: < .01 ** < .05 *

Table 6 Contingency table for multiple meta-regression moderators with chi-square results

	Learning domain		Time limitation for responses	
	Others	Text-based	Limited	Not limited / Not reported
Educational level	$\chi^2(1) = 41.76, p < .001$		$\chi^2(2) = 10.40, p = .006$	
Tertiary + Adult	127	6	56	77
Primary + Secondary	13	14	4	23
Learning domain			$\chi^2(2) = 17.79, p < .001$	
Others	–	–	56	84
Text-based learning	–	–	4	16

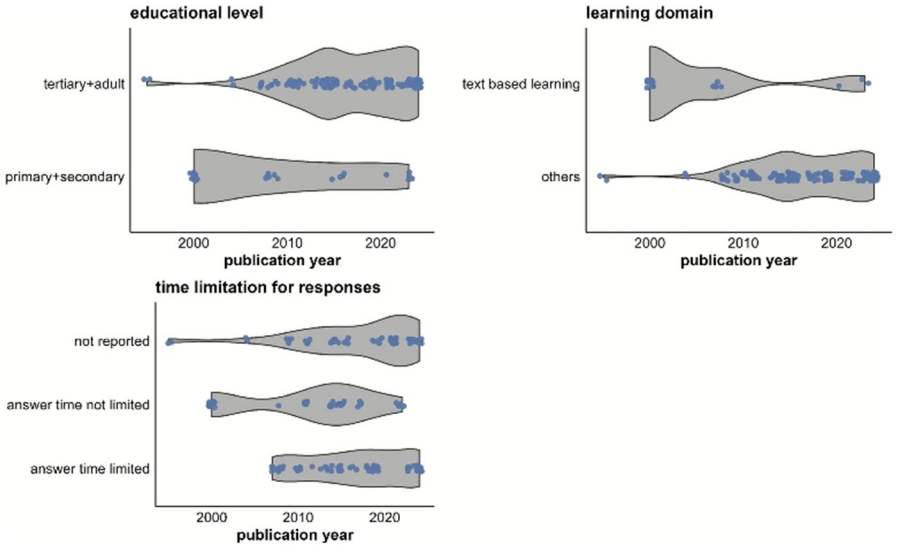


Fig. 4 Distribution of multiple meta-regression moderators across publication years

inconsistencies in the literature. Furthermore, this meta-analysis explored various moderators to explain the observed heterogeneity, including feedback characteristics (e.g., delay unit, feedback type) and study features (e.g., educational level, learning domain, and response time constraints).

Overall, the present meta-analysis found no reliable difference between immediate and delayed feedback within the range of delays represented in computer-assisted learning environments that is, from a few seconds or intervening items to several days. As summarized in Table C2, these delays span a broad continuum, with approximately half of the effects involving multi-hour or $\geq$ 10-item delays. Nevertheless, the main conclusion that feedback timing does not reliably influence learning outcomes might not generalize to substantially longer delays that are sometimes typical of classroom instruction. While the current evidence base captures the temporal dynamics most common in digital and computer-based contexts, its generalizability to extended instructional delays in authentic classroom settings remains to be established.

It is also important to note that many studies confound feedback timing with the retention interval between feedback and post-test, such that delayed feedback is often administered closer to the final test than immediate feedback. This methodological limitation, discussed in detail in the "Limitations and Directions for Future Research" section, constrains how confidently timing effects can be interpreted.

At the same time, substantial heterogeneity among studies indicates that contextual factors meaningfully influence the impact of feedback timing. Moderator analyses revealed that educational level, learning domain, and response time constraints significantly affected the direction and magnitude of timing effects, providing a partial explanation for the observed variability.

Moderator Effects in Univariate Analyses

Based on prior research, we had preregistered and systematically coded a number of moderators hypothesized to influence feedback timing effects. One such factor was the definition of immediate and delayed feedback. The literature highlights inconsistencies in how these terms are operationalized, with variations in delay units (items or time) and delay length. As a result, what is considered immediate feedback in some studies may be classified as delayed feedback in others. This inconsistency has been cited as a primary reason for mixed findings in feedback timing research (Mory, 2013; Nakata, 2015; Xu & Zeng, 2023). Quinn (2014) further argued that many studies do not introduce a sufficient temporal gap to establish a meaningful cognitive distinction between immediate and delayed feedback. Similarly, Canals et al. (2021) found no significant differences in learning outcomes, likely due to unsufficiant delay differences between conditions. To clarify this point, we categorized immediate and delayed feedback based on delay unit (items vs. time) and on the delay difference between feedback conditions. Our moderator analysis found no overall difference in outcomes based on whether delay was defined by items or time. However, the relative difference between immediate and delayed feedback—specifically in terms of the number of intervening items—significantly moderated learning outcomes, but only in studies that used item-based delay definitions. In contrast, when delay was defined in seconds or when item-based delays were converted into estimated time durations, this effect was no longer observed. These findings underscore the significant variability in how studies define and measure feedback delay, particularly when comparing item-based and time-based metrics. This inconsistency complicates the interpretation and comparison of research findings in feedback timing. Notably, the lack of a reliable correlation between the number of intervening items and the actual time elapsed before feedback delivery suggests that item-based delays can vary widely in duration across different studies. For instance, one study might implement a delay of several items that correspond to a few minutes, while another similar item-based delay could span a longer time due to differences in task complexity or participant response rates. This discrepancy indicates that item-based measures alone may not accurately reflect the temporal aspects of feedback delays. Therefore, to facilitate meaningful comparisons and syntheses across studies, it is advisable to accompany item-based delay measures with precise indications of the corresponding time durations.

Based on prior research, an important characteristic predicted to moderate the effects of feedback timing was feedback type. However, we found no evidence that the type of feedback (simple vs. elaborated) moderated the impact of feedback timing on learning outcomes. While previous meta-analyses have shown that elaborated feedback is generally more effective than simple feedback (Brummer et al., 2024; Cai et al., 2023; Wisniewski et al., 2020; Xu & Zeng, 2023), these studies focused on the main effects of feedback type rather than its interaction with timing. One exception is Taxipulati and Lu (2021), who reported that learners responded more attentively to adaptive feedback when it was provided immediately rather than after a delay—suggesting a possible interaction between timing and feedback type under specific conditions. However, they did not report whether the timing of feedback (immediate vs. delayed) differentially impacted learning outcomes when comparing

elaborated feedback to knowledge-of-correct-response feedback. Furthermore, our meta-analysis did not categorize adaptive feedback separately, as this feature was rarely reported across the included studies. The inconsistent reporting of adaptive feedback may have limited our ability to detect such interactions. Nonetheless, our findings are not necessarily at odds with the existing literature.

Educational level—another factor hypothesized to moderate the effects of feedback timing—emerged as a significant moderator in our analysis. Specifically, delayed feedback was associated with significantly better learning outcomes among primary and secondary education students, whereas no timing effect was observed in tertiary-level learners. These findings suggest that learners at different educational stages may process and benefit from feedback timing in distinct ways, possibly due to developmental differences in cognitive maturity, self-regulation, and engagement. This pattern aligns with prior claims that the impact of feedback may vary across educational levels (Swart et al., 2019). While Kulik and Kulik (1988) previously found no moderating effect of educational level, our findings may reflect shifts in educational practices and learner profiles since then—particularly in light of the increased integration of technology in instruction. Moreover, this pattern may partly reflect a temporal shift in participant profiles in more recent studies that were not included in Kulik's meta-analysis. This covariance between educational level and publication year is discussed further in Section "Moderator Effects In Multivariate Analyses".

As outlined in the introduction, we also hypothesized that the optimal timing of feedback would vary across learning domains due to differences in task characteristics and cognitive demands (Mertens et al., 2022; Van der Kleij et al., 2015). Consistent with this hypothesis, our analysis found that the learning domain significantly moderated the effect of feedback timing. Delayed feedback was more beneficial for tasks involving text memory. In contrast, domains such as cognitive skills, language learning, and STEM did not show a consistent advantage for either immediate or delayed feedback. These domain-specific effects may reflect differences in how learning unfolds across subject areas. For example, in domains like text memory, where tasks often involve recalling and organizing verbal information, delayed feedback may promote deeper processing by allowing time for retrieval and self-monitoring, consistent with theories of spacing and desirable difficulties (Bjork & Bjork, 2011; Carpenter, 2014). In contrast, domains like STEM and language learning, which were better represented in the dataset, may involve more procedural or fluency-based tasks where feedback timing plays a less consistent role. The absence of a clear timing effect in these areas may also reflect greater heterogeneity in task types and instructional designs.

Training task complexity, categorized as higher or lower order according to Bloom et al. (1964), was another preregistered moderator that showed no significant effect. Based on prior research, we predicted that immediate feedback would be more effective for lower-order tasks, whereas delayed feedback would yield greater benefits for higher-order tasks (Shute, 2008; Van der Kleij et al., 2015). However, our results did not support this expectation. One possible explanation is that our analysis focused on post-test outcomes, while training task complexity may be more relevant during the acquisition phase, warranting further investigation.

Adjustment of effect sizes on prior knowledge was also preregistered as a potential moderator, given the suggested interaction between feedback timing and learner ability (Nathan et al., 2002). However, our findings showed no significant moderating effect, suggesting that the experimental studies included in our meta-analysis effectively randomized participants across feedback timing conditions, as expected.

Lastly, we explored whether the nature of the learning task —whether designed purely for research purposes or integrated into a real curriculum — acted as a moderator. This expectation was based on Kulik and Kulik (1988), which found that applied settings generally favored immediate feedback, whereas laboratory studies tended to support delayed feedback. We hypothesized that tasks relevant to real-world applications and directly beneficial to learners' educational or professional goals might influence motivation and cognitive load, potentially affecting the efficacy of immediate versus delayed feedback differently. However, our analyses did not support this hypothesis.

Although not preregistered, we explored several additional moderators that appeared to vary systematically across studies during the screening process and that have theoretical grounding and prior support in the literature as potential explanations for inconsistencies in feedback timing effects.

One such moderator was the type of the presence or absence of response time constraints during training, which emerged as a significant moderator in our analysis. We observed that studies allowing unlimited response time tended to report stronger effects for delayed feedback compared to studies with time-limited responses. Although this variable has not been previously examined in meta-analyses on feedback timing, our findings align with cognitive load theory (Sweller, 1988; Sweller et al., 2019). Without time constraints, learners can engage in deeper cognitive processes, such as self-questioning and integrating new information, potentially enhancing the benefits of delayed feedback. In contrast, limited response time may increase cognitive load and hinder deep processing, reducing the effectiveness of delayed feedback. Notably, around 30% of studies did not specify response time constraints, indicating a gap in research reporting. Future studies should explicitly report this variable, given its significant influence on feedback effectiveness.

We also examined publication year as a potential moderator and found evidence that the effect of feedback timing has shifted over time. Earlier studies tended to report stronger benefits of delayed feedback. In contrast, more recent studies have increasingly favored immediate feedback or found no timing advantage, potentially due to the growing integration of digital technologies that enable real-time interaction, as well as shifts in learner expectations and habits. This temporal trend suggests that the effectiveness of feedback timing may be context-dependent and influenced by broader changes in educational environments and research methodologies. To further explore this possibility, we examined whether publication year correlated with other significant moderators—this is reported below in multivariate analyses of moderators.

In addition to the significant moderators above, we also explored several theoretically grounded variables that did not show significant moderating effects but warrant discussion.

One such variable was the presence of an error reminder. Our reading of the literature suggested that, in studies where immediate feedback was more effective, delayed feedback was often provided by reminding the learner's error before present-

ing the correction (e.g., Li et al. 2016). In contrast, studies using delayed feedback often did not prompt learners to recall their initial answers before or during feedback (e.g., Butler et al. 2007, Carpenter and Vul 2011). This pattern aligned with the interference-perseveration hypothesis (Kulhavy & Anderson, 1972), which suggests that error reminders may enhance the effects of immediate feedback, whereas their absence could favor delayed feedback. However, our meta-analysis found no significant moderating effect of the presence of an error reminder, thus failing to support the interference-perseveration hypothesis.

We also examined the retention interval difference between feedback and post-test as a potential moderator, given that shorter delays generally enhance memory retention (Cepeda et al., 2008; Metcalfe et al., 2009). Some studies suggest that controlling for lag-to-test intervals reduces differences in effectiveness between immediate and delayed feedback (Metcalfe et al., 2009; Nakata, 2015). However, many studies confound delayed feedback with extended retention intervals, making it unclear whether observed benefits stem from feedback timing or time elapsed before testing. Similarly, training duration may influence feedback effects, as longer training phases in delayed conditions inherently extend retention intervals, potentially impairing recall (Cepeda et al., 2008; Rohrer et al., 2005). Although our meta-analysis did not find significant moderating effects for either retention interval or training length, these null results may reflect limited statistical power due to a small number of studies that reported or controlled for these variables. Future research should report these factors more systematically to enable a clearer interpretation of their role in feedback effectiveness.

Moderator Effects In Multivariate Analyses

While individual moderators like educational level, learning domain, and response time constraints showed significant moderating effects when analyzed separately, their respective influence decreased in the multiple meta-regression. This discrepancy, coupled with the significant moderating effect of the publication year, led us to investigate potential interdependencies between these study characteristics. Indeed, our analysis revealed that these factors were not independent from each other, but rather exhibited significant covariation. For example, educational level and learning domain were closely associated, with tertiary/adult education studies primarily focusing on non-text-based learning, while primary/secondary studies were more varied.

Furthermore, an analysis of publication trends revealed systematic shifts in study characteristics over time. Experimental paradigms shifted from examining text-based learning without time constraints in primary/secondary school pupils to studies examining other types of learning domains under time-limited conditions, in tertiary/adult participants.

These dependencies suggest that it is not possible to completely disentangle the effects of these various moderators based on the existing set of studies. Rather, it is possible to characterize the studies that typically yield greater delayed vs. immediate feedback effects.

Studies that suggested an advantage of delayed feedback were typically published before 2012, involved text memorization, not limit response limit, defined feedback delay in number of items, defined delayed feedback by a large (> 21) number of

items, and were carried out on primary/secondary pupils. On the contrary, studies that suggest no difference between delayed and immediate feedback tended to be published more recently, involved learning in cognitive skills, second language or STEM and were carried out on tertiary/adult participants. Finally, studies that suggested an advantage of immediate feedback defined feedback delays in terms of a small number of items.

The overall findings of this meta-analysis partially align with previous research on feedback timing, but also help reconcile their inconsistencies by identifying patterns across study characteristics. Earlier meta-analyses, such as Azevedo and Bernard (1995), reported a clear advantage for immediate feedback, while others, like Swart et al. (2019), found benefits for delayed feedback in specific domains such as reading comprehension. More recently, reviews like Xu and Zeng (2023) have concluded that immediate feedback is often more or equally effective, depending on contextual factors.

Our results suggest that these differing conclusions reflect systematic variation in study-level characteristics rather than a universal advantage of one feedback timing over another. These patterns suggest that the effectiveness of feedback timing is highly context-dependent and has evolved alongside shifts in learning context, participant populations, and instructional design within computer-based learning environments over time.

Theoretical Implications

Our meta-analysis offers a timely opportunity to revisit the theoretical perspectives outlined in the introduction and evaluate how well their assumptions align with the available evidence.

The interference-perseveration hypothesis posits that delayed feedback enhances error correction by allowing incorrect responses to fade from working memory, thereby reducing interference. However, our results provide little support for this mechanism. Most studies did not report whether feedback effects differed based on the accuracy of initial responses, nor did they systematically manipulate whether errors were reactivated during feedback. Although we examined whether the presence of answer reminders moderated feedback effects and found no significant interaction, this analysis relied on a coded feature indicating the presence of reminders rather than on experimental manipulation. One notable exception is Iwaki et al. (2017), who found that reminding learners of their errors did not impair learning, but it improved performance, undermining the core assumption of the interference account. Future studies should condition outcomes on initial correctness and directly manipulate error reactivation to more rigorously test this hypothesis.

In contrast to error-based theories, memory-based frameworks such as the distributed practice account argue that delayed feedback is beneficial not because of error fading, but because it introduces spacing between initial response and corrective feedback, effectively creating a second retrieval opportunity. Although this account aligns with robust evidence on the spacing effect (Cepeda et al., 2008), few studies in our meta-analysis directly tested this mechanism (e.g., Smith and Kimball 2010). Furthermore, delay difference did not emerge as a significant moderator, indicating that longer delays did not consistently lead to better outcomes, even when

long enough to plausibly induce spacing effects. Since spacing is theorized to benefit primarily initially correct responses, future studies should report outcomes separately by initial response accuracy and ensure that retrieval opportunities and inter-study intervals are held constant. The study by Smith and Kimball (2010) exemplifies how such methodological controls can be implemented to rigorously evaluate spacing-based predictions in the context of feedback timing.

The recursive reminding framework suggests that delayed feedback supports learning by reactivating the learner's original response, allowing past and current information to be integrated into a unified memory trace. However, our meta-analysis found no significant moderating effect of whether feedback prompted recall of the initial answer. At the same time, few studies directly tested whether such prompts successfully triggered reactivation, or whether initial responses were truly forgotten in studies without reminders. Therefore, while our findings challenge a core assumption of the recursive reminding account, more targeted experiments are needed to clarify whether and when reactivation occurs and whether it contributes to the learning benefits of delayed feedback.

Thus, most studies included in this meta-analysis fall short of testing core assumptions related to error reactivation, spacing, and initial response accuracy. This limitation highlights the need for future research using theory-driven designs that directly examine these mechanisms to clarify when and why feedback timing influences learning. More broadly, the diversity of findings across contexts suggests that a more comprehensive theoretical framework may be needed: one that integrates cognitive, contextual, and learner-related factors to explain when delayed or immediate feedback is most effective.

Practical Implications

This meta-analysis has some implications for the design and implementation of educational technologies, particularly in computer-assisted learning environments where feedback timing can be precisely controlled to improve learning.

The multiple meta-regression showed that none of the significant moderators (educational level, learning domain, and time constraints) retained a unique predictive effect when considered together. This finding suggests that these factors are interrelated and reflect broader contextual patterns rather than independent influences. Consequently, the generalizability of each moderator in isolation is limited, and the effectiveness of feedback timing should be interpreted in relation to specific combinations of learner, task, and contextual characteristics.

Nonetheless, several consistent tendencies were observed. Studies involving text-based learning or conducted in primary and secondary education settings more often favored delayed feedback, whereas studies conducted in tertiary and adult learning contexts tended to show little or no benefit of delay. These patterns provide cautious but informative guidance: the effectiveness of feedback timing appears to depend on how it aligns with the cognitive demands and pacing of the learning environment.

As educational technologies continue to evolve, these insights can inform more flexible and context-sensitive approaches to feedback timing, ensuring that it supports learners' needs within appropriate pedagogical frameworks.

Limitations and Directions for Future Research

Our study is subject to several limitations. One primary limitation is the substantial heterogeneity observed across studies, which was not adequately accounted for by most moderator analyses. A key contributing factor was the inconsistent reporting of critical study variables—particularly among categorical moderators—which constrained our ability to conduct more comprehensive meta-regressions. Additionally, some studies lacked sufficient data to compute effect sizes despite meeting other inclusion criteria. As a result, the small number of effect sizes reduced the reliability of certain findings, making it difficult to determine whether observed effects were robust or a result of limited data. This issue was particularly evident in comparisons of feedback timing effects across educational levels. The majority of studies focused on university students, with relatively few examining primary education ($n = 5$), limiting the ability to draw conclusions about younger learners. Similarly, certain learning domains, such as social sciences ($n = 3$) and reading comprehension ($n = 2$), were underrepresented, reducing the generalizability of findings. With such limited power, we can only conclude that additional studies are needed to improve the reliability of the findings (Harrer et al., 2021).

Another limitation concerns the inconsistent conceptualization of feedback timing in the literature. To accommodate this variation, we categorized feedback delays into time-based and item-based groups. This division, however, reduced statistical power. To regain power, we attempted to approximate item-based delays in temporal terms using additional study information. Unfortunately, these details were not consistently reported, making such conversions unreliable. Thus, more experimental studies are needed to clarify inconsistencies in feedback delay definitions and their effects on learning outcomes since meta-analyses cannot generate new findings beyond what is available in the experimental literature.

Another critical methodological limitation in the literature and, by extension, in this meta-analysis is the frequent confounding of feedback timing with retention interval. Except for eight studies (Butler et al., 2007; Canals et al., 2021; Hays et al., 2013; Lu et al., 2021; Metcalfe et al., 2009; Nakata, 2015; Smith & Kimball, 2010; Tanaka et al., 2019), most of the included studies did not experimentally control that the time between the last feedback and the post-test was equivalent across immediate and delayed feedback conditions. As a result, delayed feedback was often administered closer to the test than immediate feedback, potentially inflating its apparent effectiveness due to recency effects. This concern has been empirically validated in the literature. Metcalfe et al. (2009) demonstrated that delayed feedback's benefits can often be attributed to its shorter interval to the test. Similarly, Nakata (2015) showed that when this interval is equated, timing effects tend to disappear. Although our coding captured this issue, including the retention interval difference as a moderator, our meta-analysis did not find a significant moderating effect. This suggests that, at the aggregate level, the effectiveness of feedback timing may not strongly depend on proximity to the test. However, this result must be interpreted cautiously, as very few studies explicitly controlled this interval, limiting the power and interpretability of our moderator analysis.

Disentangling these factors requires experimental designs that independently vary both feedback and test intervals in a factorial design. Drawing inspiration from methodologies used in spacing effect research can guide future work on this. Spacing effect research, like studies on feedback timing, faces similar constraints related to the interval between training and the post-test. Among these, the factorial designs employed by Cepeda et al. (2008) offer a particularly relevant model. An analogous study could systematically manipulate both feedback timing and retention interval to examine how these variables interact. This design would enable researchers to examine the interaction between these two temporal factors and to identify the conditions under which feedback timing is better. More importantly, it would help determine whether any observed benefit of delayed feedback reflects a genuine cognitive advantage or merely results from its proximity to the test.

A further limitation relates to the range of moderators examined. Prior research suggests that various feedback characteristics, study features, and learner attributes may influence the effectiveness of feedback, but many of these could not be systematically coded due to insufficient reporting. For instance, Clariana et al. (2000) suggested that learning material difficulty may influence optimal feedback timing. Yet, comparing difficulty levels across studies is inherently challenging, as they are often not measured on a common scale and are operationalized differently depending on the context. Furthermore, Timmers and Veldkamp (2011) proposed that learners' prior knowledge could moderate feedback effectiveness, while Nathan et al. (2002) highlighted the interaction between feedback timing and learner ability. However, as difficulty level and learner ability were not systematically reported, moderator analyses on these factors were not feasible. Future research should investigate these moderators in both primary studies and meta-analyses.

Furthermore, the findings of this meta-analysis are based primarily on controlled experimental settings, particularly in computer-assisted tutoring systems, where feedback delays are precisely controlled and typically shorter than those encountered in real classroom contexts. Most studies measured effects in the short term, using post-tests administered immediately or shortly after feedback interventions. This approach may not fully capture real-world educational complexities, where factors such as classroom dynamics, student motivation, and instructor feedback practices play a role. Thus, the ecological validity of our results is limited to this setting, and future experimental research should directly examine feedback timing in classroom settings that employ week-long feedback intervals.

Finally, this meta-analysis focused solely on feedback timing's impact on learning outcomes, as measured by post-test scores. Future research should also examine its effects on student engagement, as studies suggest that immediate feedback is often preferred by educators and students (Lefevre & Cox, 2017; Mullet et al., 2014; Van der Kleij et al., 2012). Investigating the balance between learning effectiveness and engagement would provide a more comprehensive understanding of feedback timing in educational contexts.

Given that our multivariate moderator analysis was limited by the fact that moderators were partly confounded with each other, providing more definitive conclusions about the respective role of each moderator would require new studies with specific characteristics that were missing in the current literature. In particular, it

would be useful to have more studies comparing immediate and delayed feedback: 1) in primary and secondary school pupils in general; 2) involving text-based learning in tertiary/adult participants; 3) involving limited response time in primary/secondary school participants; 4) involving text-based learning and limited response time.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-026-10117-8.

Acknowledgements This work was funded by Agence Nationale de la Recherche, grants ANR-17-EURE-0017, ANR-10-IDEX-0001-02 and ANR-21-CE28-0030

Data Availability The data and analytic code necessary to reproduce the analyses presented in this paper are publicly accessible: https://github.com/ervanihank/Meta_Analysis_Feedback_Timing.git. The analyses presented here were preregistered: https://osf.io/nf84e?view_only=9342a041d11b4bc7809f92383c4a40d4.

Declarations

Conflicts of Interest none.

References

Attali, Y., & van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education, 110*, 154–169.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*(2), 111–127.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society, 2*(59–68), 56–64.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1964). *Taxonomy of educational objectives* (Vol. 2). Longmans, Green New York.

Brummer, L., de Boer, H., Mouw, J. M., & Strijbos, J.-I. W. (2024). A meta-analysis of the effects of context, content, and task factors of digitally delivered instructional feedback on learning performance. *Learning Environments Research*, 1–24,

Butler, A. C., & Woodward, N. R. (2018). Toward consilience in the use of task-level feedback to promote learning. *Psychology of learning and motivation* (Vol. 69, pp. 1–38). Elsevier.

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin & Review, 18*, 1238–1244.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604–616.

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

Cai, Z., Gui, Y., Mao, P., Wang, Z., Hao, X., Fan, X., & Tai, R. H. (2023). The effect of feedback on academic achievement in technology-rich learning environments (tres): A meta-analytic review. *Educational Research Review, 39*, 100521.

Canals, L., Granena, G., Yilmaz, Y., & Malicka, A. (2021). The relative effectiveness of immediate and delayed corrective feedback in video-based computer-mediated communication. *Language Teaching Research*, 13621688211052793

Carpenter, S. K. (2014). Spacing and interleaving of study and practice. *Applying the science of learning in education: Infusing psychological science into the curriculum*, 131–141.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24*, 369–378.

Carpenter, S. K., & Vul, E. (2011). Delaying feedback by three seconds benefits retention of face-name pairs: The role of active anticipatory processing. *Memory & Cognition, 39*, 1211–1221.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095–1102.

Clariana, R. B., Wagner, D., & Roher Murphy, L. C. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development, 48*(3), 5–22.

Corral, D., Carpenter, S. K., & Clingan-Siverly, S. (2021). The effects of immediate versus delayed feedback on complex concept learning. *Quarterly Journal of Experimental Psychology, 74*(4), 786–799.

Dempsey, J. V., & Wager, S. U. (1988). A taxonomy for the timing of feedback in computer-based instruction. *Educational Technology, 28*(10), 20–25.

Donkin, R., Askew, E., & Stevenson, H. (2019). Video feedback and e-learning enhances laboratory skills and engagement in medical laboratory science students. *BMC Medical Education, 19*, 1–12.

Dunbar, N. E., Jensen, M. L., Miller, C. H., Bessarabova, E., Lee, Y.-I.H., Wilson, S. N., . . . others (2017). Mitigation of cognitive bias with a serious game: Two experiments testing feedback timing and source. *International Journal of Game-Based Learning (IJGBL), 7*(4), 86–100.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj, 315*(7109), 629–634.

Fisher, Z., & Tipton, E. (2015). robumeta: An r-package for robust variance estimation in meta-analysis. arXiv:1503.02220

Foerde, K., & Shohamy, D. (2011). Feedback timing modulates brain systems for learning in humans. *Journal of Neuroscience, 31*(37), 13157–13167.

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. (2021). *Doing meta-analysis with r: A hands-on guide*. Chapman and Hall/CRC.

Hattie, J., & Clarke, S. (2018). *Visible learning: feedback*. Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 290.

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.

Hull, C. L. (1952). *A behavior system; an introduction to behavior theory concerning the individual organism*. Yale University Press.

Iwaki, N., Nara, T., & Tanaka, S. (2017). Does delayed corrective feedback enhance acquisition of correct information? *Acta Psychologica, 181*, 75–81.

Jaehnig, W., & Miller, M. L. (2007). Feedback types in programmed instruction: A systematic review. *The Psychological Record, 57*(2), 219–232.

Janelli, M., & Lipnevich, A. A. (2021). Effects of pre-tests and feedback on performance outcomes and persistence in massive open online courses. *Computers & Education, 161*, 104076.

Kmet, L. (2004). Standard quality assessment criteria for evaluating primary research papers from a variety of fields. *Alberta Heritage Foundation for Medical Research Edmonton*.

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 63*(5), 505.

Kulik, J. A., & Kulik, C.-L.C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79–97.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374.

Latimier, A., Peyre, H., & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educational Psychology Review, 33*, 959–987.

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language, Learning & Technology*, *19*(2).

Lefevre, D., & Cox, B. (2017). Delayed instructional feedback may be more effective, but is this contrary to learners' preferences? *British Journal of Educational Technology, 48*(6), 1357–1367.

Lim, L.-A., Gentili, S., Pardo, A., Kovanović, V., Whitelock-Wainwright, A., Gašević, D., & Dawson, S. (2021). What changes, and for whom? A study of the impact of learning analytics-based process feedback in a large course. *Learning and Instruction, 72*, 101202.

Lipsey, M. W. (2001). Practical meta-analysis. *Thousand Oaks*

Li, S., Zhu, Y., & Ellis, R. (2016). The effects of the timing of corrective feedback on the acquisition of a new linguistic structure. *The Modern Language Journal, 100*(1), 276–295.

Lu, X., Sales, A., & Heffernan, N. T. (2021). Immediate versus delayed feedback on learning: Do people's instincts really conflict with reality? *Journal of Higher Education Theory and Practice*, *21*(16).

Mason, B.J., & Bruning, R. (2001). Providing feedback in computer-based instruction: What the research tells us. *Retrieved February*, *15*, 2007.

Mertens, U., Finn, B., & Lindner, M. A. (2022). Effects of computer-based feedback on lower-and higher-order learning outcomes: A network meta-analysis. *Journal of Educational Psychology, 114*(8), 1743.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37*(8), 1077–1087.

Metcalfe, J., & Xu, J. (2018). Learning from one's own errors and those of others. *Psychonomic Bulletin & Review, 25*, 402–408.

Mohamed, S., & Adnan, F. (2020). Feedback in computer-assisted language learning: A meta-analysis. *Tesl-Ej, 24*(2), Article n2.

Mohsen, M. A. (2022). Computer-mediated corrective feedback to improve l2 writing skills: A meta-analysis. *Journal of Educational Computing Research, 60*(5), 1253–1276.

Mory, E. H. (2013). Feedback research revisited. *Handbook of research on educational communications and technology* (pp. 738–776). Routledge.

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition, 3*(3), 222–229.

Nakata, T. (2015). Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning? *Language Teaching Research, 19*(4), 416–434.

Nathan, M. J., Stephens, A. C., Masarik, D., Alibali, M. W., & Koedinger, K. R. (2002). Representational fluency in middle school: A classroom study. *Proceedings of the twenty-fourth annual meeting of the North American chapter of the international group for the psychology of mathematics education* (Vol. 1, pp. 462–472).

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D. ... others (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, *372*.

Peeck, J., Van den Bosch, A., & Kreupeling, W. (1985). Effects of informative feedback in relation to retention of initial responses. *Contemporary Educational Psychology, 10*(4), 303–313.

Peeck, J., & Tillema, H. (1979). Learning from feedback: Comparison of two feedback procedures in a classroom setting. *Perceptual and Motor Skills, 48*(2), 351–354.

Phye, G. D., & Andre, T. (1989). Delayed retention effect: Attention, perseveration, or both? *Contemporary Educational Psychology, 14*(2), 173–185.

Quinn, P. (2014). *Delayed versus immediate corrective feedback on orally produced passive errors in English* (Unpublished doctoral dissertation). University of Toronto Toronto.

Rohrer, D., Taylor, K., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology, 19*(3), 361–374.

Roper, W. (1977). Feedback in computer assisted instruction. *Programmed Learning and Educational Technology, 14*(1), 43–49.

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics, 17*(4), 363–374.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Saltzman, I. J. (1951). Delay of reward and human verbal learning. *Journal of Experimental Psychology, 41*(6), 437.

Schroth, M. L. (1992). The effects of delay of feedback on a delayed concept formation transfer task. *Contemporary Educational Psychology, 17*(1), 78–82.

Schroth, M. L. (1995). Variable delay of feedback procedures and subsequent concept formation transfer. *The Journal of General Psychology, 122*(4), 393–399.

Schroth, M. L., & Lund, E. (1993). Role of delay of feedback on subsequent pattern recognition transfer tasks. *Contemporary Educational Psychology, 18*(1), 15–22.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Skinner, B. F. (1965). *Science and human behavior* (No. 92904). Simon and Schuster.

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 80.

Smits, M. H., Boon, J., Sluijsmans, D. M., & Van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments, 16*(2), 183–193.
Surber, J. R., & Anderson, R. C. (1975). Delay-retention effect in natural classroom settings. *Journal of Educational Psychology, 67*(2), 170.
Swart, E. K., Nielen, T. M., & Sikkema-de Jong, M. T. (2019). Supporting learning from text: A meta-analysis on the timing and content of effective feedback. *Educational Research Review, 28*, 100296.
Sweller, J. (2011). Cognitive load theory. *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Elsevier.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.
Sweller, J., Van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*, 261–292.
Tanaka, S., Miyatani, M., & Iwaki, N. (2019). Response format, not semantic activation, influences the failed retrieval effect. *Frontiers in Psychology, 10*, 599.
Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and spss. *Research Synthesis Methods, 5*(1), 13–30.
Taxipulati, S., & Lu, H.-I.D. (2021). The influence of feedback content and feedback time on multimedia learning achievement of college students and its mechanism. *Frontiers in Psychology, 12*, 706821.
Timmers, C., & Veldkamp, B. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education, 56*(3), 923–930.
Van der Kleij, F. M., Eggen, T. J., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education, 58*(1), 263–272.
Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511.
Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 487662.
Xu, M., & Zeng, S. (2023). Optimal timing of treatment for errors in second language learning-a systematic review of corrective feedback timing. *Frontiers in Psychology, 14*, 1026174.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Erva Nihan Kandemir[1] · Emanuele Esposito[1] · Lilas Gurgand[1] · Franck Ramus[1]

✉ Erva Nihan Kandemir
erva.nihan.kandemir@ens.psl.eu

[1] Laboratoire De Sciences Cognitives Et Psycholinguistique, Département d'Etudes Cognitives, Ecole Normale Supérieure, CNRS, PSL University, 29 rue d'Ulm, 75005 Paris, France