**Does Computer-Based Feedback Foster Expository Writing?**

**A Meta-Analysis**

# Abstract

Feedback is a beneficial instructional strategy to improve students’ writing. To this end, computer-based feedback technologies are discussed as promising supplements in addition to instructor feedback. Several computer-based feedback approaches have been developed. However, to date, if and under which conditions computer-based feedback can be effective is unclear. We conducted a meta-analysis to investigate the effectiveness of computer-based feedback on writing quality, and to explore its boundary conditions (*k* = 24 studies with pre-posttest design; *k* = 15 studies with pretest-posttest-control group design). The results suggest that computer-based feedback contributed to writing quality with a small to medium effect (*g* = .53), but only for studies with pretest-posttest design. For studies with pretest-posttest-control group design, the effect (*g* = .27) was not significant. Moderator analyses revealed that firstly text level was a significant moderator: Computer-based feedback was only effective for enhancing surface features of the text (e.g., grammar, spelling), as such feedback was likely easier to implement. Secondly, feedback specificity was a strong moderator as specific computer-based feedback improved students’ text quality more than generic feedback. The results highlight the need for the development of sophisticated technologies to provide students with high-level feedback and subsequently scaffold their expository writing.

*Keywords:* computer-based feedback, writing quality, meta-analysis, technology-enhanced learning and instruction

# Introduction

Writing comprehensible and elaborated expository texts is regarded as a key 21st-century skill (Concha & Paratore, 2010; Geiser & Studley, 2001; Lachner et al., 2017a; Powell, 2009; Sharp, 2016). When producing comprehensible texts, students need to iteratively plan, execute, and revise their writing (e.g., Hayes & Flower, 1986; Kellogg & Whiteford, 2009), which puts high demands on students. Feedback has been demonstrated as a beneficial instructional strategy to particularly support students’ writing processes, as it enhances processes of monitoring and regulation (Graham et al., 2015; Kellogg & Raulerson, 2007; Kieft et al., 2008; Philipp, 2017; Roscoe & McNamara, 2013). Although feedback has been demonstrated to be effective, it is regarded to be rather time-consuming and laborious (Brindle et al., 2015; Graham, 2019; Graham & Hebert, 2011). To this end, computer-based feedback has been discussed as a potential supplement to support students’ writing (Kellogg & Whiteford, 2009; Roscoe & McNamara, 2013; Strobl et al., 2019). To date, several feedback technologies have been developed. However, meta-analytic evidence is missing, that allows to draw conclusions regarding whether and under what conditions computer-based feedback is generally effective for enhancing students’ writing.

To close this research gap, we conducted a meta-analysis to 1) estimate the overall effectiveness of computer-based feedback in the context of expository writing, 2) as well as examining potential boundary conditions. Therefore, we particularly focused on boundary conditions which may have resulted from system-related design aspects of the computer-based feedback (i.e., the specificity, the format of representation, or the type of text level that was addressed within the feedback), as well as individual boundary conditions, based on students’ prerequisites (i.e., prior knowledge).

## Expository Writing

Writing can be regarded as a self-regulated process (e.g., Harris et al., 2008; Hayes, 2012; Hayes & Flower, 1986; Philipp, 2014). Hayes (2012) proposed three different self-regulation levels which constitute effective writing: (1) The *control level* comprises a writer’s motivation including the goals and plans of the subsequent writing process and the product, as well as her or his rhetorical and content-related prior knowledge (Becker-Mrotzek et al., 2014; Bereiter & Scardamalia, 1987). (2) The *process level* comprises different sub-processes to achieve the writing goals: First, a writer proposes a writing plan, translates the proposed ideas into propositions and subsequently transcribes them into a written draft (Hayes & Olinghouse, 2015). Based on the draft the writer ideally iteratively monitors and revises the draft to create a final writing product which ideally attains the previously set writing goals. (3) The *resource level* consists of generic resources, such as attention, working memory resources, and pre-requisite skills, such as reading.

The model by Hayes (2012) illustrates that writing is a complex task, requiring assistance during early stages of writing (Becker-Mrotzek et al., 2014; Bereiter & Scardamalia, 1987; Flower & Hayes, 1980; Hayes & Flower, 1986; Kellogg & Whiteford, 2009). The need for additional support, particularly holds true for revision processes, as seminal studies indicated that – compared to experienced writers – novice writers less frequently engage in revision activities that tackle substantive aspects of their writing (see Chanquoy, 2009; Cho & MacArthur, 2010; Lachner et al., 2017b; MacArthur & Graham, 2016; Roscoe et al., 2016).

## Feedback to Support Revision Processes

Formative feedback is one of the most effective strategies for stimulating students’ revision activities and enhance writing quality (Graham et al., 2015; Philipp, 2017). Commonly, formative feedback is understood as information about distinct aspects of a learner’s current performance (e.g., writing quality). The goal of feedback is to provide information to help the recipient reduce discrepancies between the actual and the targeted performance (Hattie & Timperley, 2007; Narciss, 2008; Stevenson & Phakiti, 2014; Winne & Butler, 1994). Formative feedback, thus, not only includes assessment aspects, but also information on how to regulate the further learning process and improve their current performance (i.e., quality of the writing product) regarding assessment criteria (Butler & Winne, 1995; Narciss, 2008).

Recent meta-analyses showed that formative feedback can support students in developing their writing skills. For example, Graham et al. (2015) examined effects of generic formative assessment and feedback on children’s (kindergarten to grade 8) writing quality based on 16 empirical studies. The authors obtained a medium to large effect (*d* = 0.77) of formative feedback on writing quality compared to another treatment condition or a control group that did not receive feedback (see also Wisniewski et al., 2020, for related findings on the general effectiveness of feedback).

## Computer-Based Feedback – A Feasible Approach to Foster Expository Writing?

Given that formative feedback is often regarded as time-consuming and laborious, several computer-based feedback technologies have been proposed as a supplement to instructor feedback (e.g., Allen et al., 2016; Kellogg & Whiteford, 2009; Roscoe & McNamara, 2013). For instance, automated writing evaluation systems (AWE), such as Criterion, use natural language processing (NLP) technologies, to provide feedback to students and initiate distinct revision activities (Dikli, 2006; Shermis & Burstein, 2003; Wilson & Roscoe, 2020; see Lachner et al., 2017a; Roscoe & McNamara, 2013, for other examples of tools using NLP). AWE-systems, therefore, synthesize feedback on several dimensions, ranging from low-level features, such as grammar or spelling to high-level features, such as the cohesion or the complexity of a text. In most cases, the AWE system generates a holistic score across these dimensions, which is then automatically provided as feedback to the learner. This feedback is represented in different formats, such as numeric, text-based or graph-based information.

Similarly, other feedback systems, for instance, Intelligent Tutoring Systems (ITS; e.g., Summary Street, Writing Pal) use algorithms which compare students’ drafts with writing samples which were assessed by writing experts. ITS also combine different representation formats of feedback, e.g., written comments/text-based feedback and additional scores when providing feedback to learners. Yet, other feedback tools (e.g., CohViz) use graphical approaches, based on network analysis and decision trees to provide graphical feedback regarding the cohesion of students’ texts (Burkhart et al., 2020).

Strobl et al. (2019) conducted a systematic review to map the research landscape on computer-based feedback tools for writing support. The authors showed that there is a wide variety of existing tools that differ in terms of system-related aspects of the feedback (e.g., specificity of the provided information or the targeted text level), but also regarding the learner-related settings (e.g., knowledge level, educational level), in which the feedback was implemented. Most of the other previous research syntheses investigating computer-based feedback on writing focused exclusively on AWE systems. For instance, Nunes et al. (2021) and Fleckenstein, Reble et al. (2023) conducted systematic reviews on AWE feedback and identified potential influencing factors (e.g., role of teachers, integration of AWE systems in instructional programs) and research gaps (e.g., lack of studies with experimental control group designs, little consideration of individual learner-related influencing factors). However, they were not able to quantify the effect of AWE feedback on writing.

To this end, however, meta-analytical evidence regarding the effectiveness of computer-based feedback in the context of writing quality is scarce. Initial evidence can be found in the subgroup analyses of the meta-analysis by Graham et al. (2015). In this subgroup analysis, the authors synthesized the effects of computer-based feedback approaches on writing quality, based on *k* = 4 studies. The authors obtained a small to medium (*d* = .38, 95% CI [.17, .59]) significant effect with a large heterogeneity. The restricted number of included studies, however, requires a careful interpretation of the findings. Furthermore, these studies differed from those which were included in the meta-analytic report by Graham et al. (2011). The overall effect over the studies (*k* = 4) in Graham et al. (2011) was not significant, indicating that the effect of computer-based feedback is not robust and further research is needed.

More recently, Fleckenstein, Liebenow and Meyer (2023), Ngo et al. (2022), and Zhai and Ma (2021) performed meta-analyses to further examine the effectiveness of AWE systems. Each of these meta-analyses revealed moderate to large (*g* = 0.55 to *g* = 0.86) positive effects of AWE feedback on learners’ overall writing quality. However, besides the fact that these meta-analyses focused only on AWE tools, all of them showed a large heterogeneity across the studies (e.g., writing in first and second language with different emphases), making interpretation of the data difficult. Fleckenstein, Liebenow and Meyer (2023) found some significant moderators of the feedback effect (e.g., educational level, language, time). However, these moderator variables were more related to the intervention itself or the context. To date, it is still unclear what the feedback- and learner-related boundary conditions of computer-based feedback in general (not exclusively AWE systems) are and what factors might explain the heterogeneity found in all of the previous research syntheses.

## Potential Boundary Conditions of Computer-Based Feedback

Based on previous theoretical considerations, it can be argued that the effectiveness of computer-based feedback may depend on distinct system-related, but also learner-related boundary conditions (see Panadero & Lipnevich, 2022; Shute, 2008; Wisniewski et al., 2020, for characteristics of feedback in general; see Fleckenstein, Liebenow & Meyer, 2023; Kluger & DeNisi, 1996; Narciss, 2008; Ngo et al., 2022; Patchan et al., 2016; Strobl et al., 2019, in particular for characteristics of computer-based feedback). In the current meta-analysis, we considered four influential factors which sought to impact the effectiveness of computer-based feedback to improve students’ writing quality: specificity of feedback, representation of feedback, text level on which feedback was provided, and writers’ individual prior knowledge.

### Does the Specificity of Feedback Matter?

Computer-based feedback technologies may have the potential to provide specific information to enhance writing quality (Goodman et al., 2004). Specificity of feedback refers to the level of detail of the feedback information (Goodman et al., 2004; Patchan et al., 2016).

Previous research has shown that specific feedback contributes to enhancements of writing quality (Hattie & Timperley, 2007; Shute, 2008; but also, Wisniewski et al., 2020). To this end generic feedback is perceived as less easy to be implemented and thus may result in lower quality gains (Kluger & DeNisi, 1996; Patchan et al., 2016; Shute, 2008). For instance, Lachner et al. (2017a), compared effects of generic versus specific computer-based feedback on the cohesion of university students’ expository texts. They varied the type of feedback (specific concept map feedback vs. specific outline feedback vs. general feedback) and examined its effect on the local and global cohesion of student revisions. The results showed that writers perceived generic feedback as more difficult than specific feedback. Specific feedback also contributed to writing performance (as measured by the level of local cohesion).

### Does the Representation Format of Feedback Matter?

Previous computer-based feedback technologies not only differed regarding the specificity of the feedback implementation, but also regarding to the utilized representation. Numerical representations provide information in the form of numbers (e.g., overall scores, grades, percentages, or points achieved; Kellogg et al., 2010; Roscoe et al., 2013; Wilson & Roscoe, 2020). Numerical feedback is often visualized in the form of charts, diagrams, and rating scales. Additionally, there are rather complex embedded formats that represent feedback information directly within the text, for example, by signaling, markings, or highlighting (Burkhart et al., 2020; Moore & MacArthur, 2016; Tsai et al., 2020), allowing for easier integration of feedback into the draft. More recent approaches additionally allow for pictorial representations of a text in which the feedback information is represented, for example, as a graph, in form of a concept map, or using symbols (Lachner et al., 2017a, 2017b; O’Rourke et al., 2011; Pirnay-Dummer & Ifenthaler, 2011; Villalon & Calvo, 2011). However, it is still unclear which representation format is best suitable for providing computer-based feedback to support learners in revising their texts and to enhance their writing quality effectively. In our meta-analysis, we were interested in exploring if it makes a difference how the computer-based feedback is presented. Furthermore, we were interested in whether the combination of multiple representations would additionally contribute to writing quality than mono-representational formats (e.g., Ainsworth, 2006) as different representations can complement, constrain or relate to each other, enhancing students to process their revisions more deeply and gain a deeper understanding of the feedback (cf. Ainsworth, 2006; Scaife & Rogers, 1996).

For instance, in their experiment, Burkhart et al. (2020) compared effects of different versions of computer-based feedback on writing quality. In two conditions, students received concept map feedback which highlighted cohesion gaps of their texts in addition to their text representation. An amended version, additionally, explicitly signaled text-picture relations between the students’ draft and the concept map feedback. The contiguous feedback condition directly highlighted the cohesion gaps within the text representation (i.e., mono-representation), whereas in the no-feedback condition, students did not receive any feedback. The authors found benefits of the mono-representation for enhancing the local cohesion of a text. The amended multiple representation format (concept map and text representation) had benefits regarding global cohesion. These findings suggest that mono- versus multiple representational feedback differently contributed to students’ writing quality.

### ***Does it Matter on Which Level of Text Quality Feedback is Provided?***

Besides the design of computer-based feedback, particularly the level of text quality, that is the type of information which is provided, is considered to be important to enhance students’ writing quality (Patchan et al., 2016). Feedback on a *lower hierarchical level* (also called micro-level or low-level feedback; Strobl et al., 2019) focuses on surface features of the text and refers, for example, to grammar, spelling, text length, or word count. Feedback on a *higher hierarchical level* (also called macro-level or high-level feedback; cf. Strobl et al., 2019) refers to the organization and structure, the style, or the cohesion of the text. Commonly, it is assumed that improvements of writing quality may particularly emerge, when higher level feedback is provided (Chanquoy, 2009; Patchan et al., 2016). For instance, Patchan et al. (2016) analyzed the effects of the text-level of computer-mediated peer-comments on writing quality. The authors found that only higher-level comments contributed to writing quality. Given that most computer-based feedback systems rather provide information on a lower text level (see Strobl et al., 2019), it is an open issue, whether the text level may account for writing quality in the context of computer-based feedback.

### Does Students’ Prior Knowledge Matter?

Additionally, the level of students’ prior knowledge may affect the effectiveness of computer-based feedback. Such effects are often explained against the expertise-reversal effect, which postulates that the effectiveness of instructional interventions, such as feedback may depend on the level of prior knowledge (Bromme et al., 2004; Kalyuga, 2007; Nihalani et al., 2011). Whereas low prior-knowledge students may require additional assistance during learning, instructional support may have negative consequences for more experienced students due to redundant processing of information (Ngo et al., 2022; Richter et al., 2016).

For instance, Fyfe and Rittle-Johnson (2016) gave elementary students strategy instruction about solving mathematical equivalence problems or not to experimentally induce students’ prior knowledge. Afterwards, in a practice phase, the students solved 12 mathematical problems. Additionally, students were either provided with feedback or not. The authors found a knowledge-by-feedback interaction, as only low-prior knowledge students profited from the feedback. Whether these findings may replicate in writing settings, is an open question.

## Overview of the Present Meta-Analysis

The aims of the present study were to examine the effectiveness of computer-based feedback regarding the quality of writing, and analyze potential boundary conditions, based on the previous considerations. In contrast to earlier research syntheses, we focused on formative feedback (excluding summative feedback) and system-generated feedback only. Thus, we did not include computer-mediated feedback (e.g., via peer-review systems) which could likely confound our findings. Additionally, we focused on first language writing.

Previous meta-analyses that examined the effectiveness of computer-based feedback on writing could only include a restricted set of studies (Graham et al., 2015; considering articles from 1975 to 2011) or focused on AWE systems only (Fleckenstein, Liebenow, & Meyer, 2023; Nog et al., 2022; Zhai & Ma, 2022). Therefore, we aimed to update the meta-analytical knowledge, and additionally investigated potential feedback-related and student-related boundary conditions for the effective use of different computer-based feedback technologies to support writing.

# Method

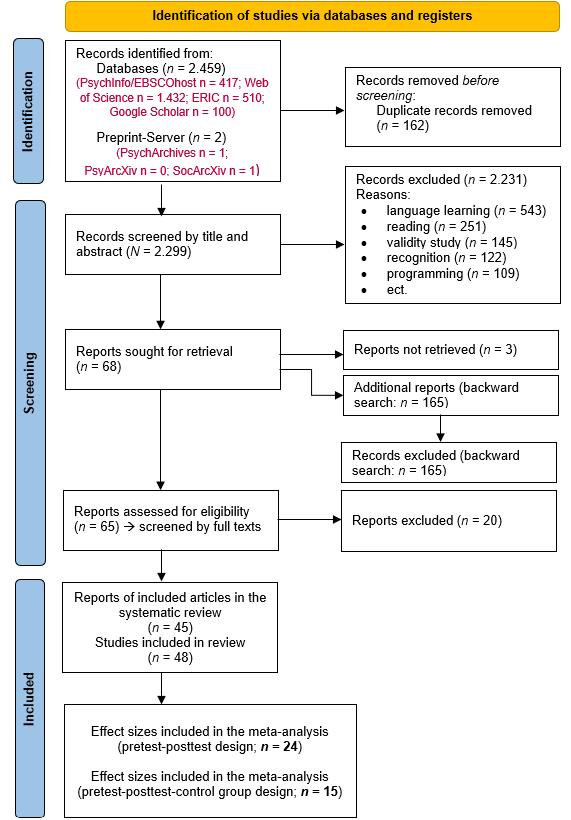
Based on Borenstein et al. (2009), we conducted the meta-analysis according to the three proposed steps: Identifying relevant studies, coding study characteristics, and computing mean effect sizes and their statistical significance. The meta-analysis was preregistered (see [preregistration](https://unitc-my.sharepoint.com/personal/qxowa01_cloud_uni-tuebingen_de/Documents/Metaanalyse/Manuskript/preregistration)) via *preregRS* (Schneider et al., 2021).

## Identifying Relevant Studies

Using the PICO framework (Population – Intervention – Comparison – Outcome) framework (Fineout-Overholt & Johnston, 2005; Schardt et al., 2007), we developed the following search term: `((computer\* OR automat\*) AND (writ\* AND (argument\* OR essay\* OR summary OR exposit\* OR expla\*)) AND (feedback OR evaluat\* OR assess\* OR scor\*) NOT (peer OR medic\* OR neural-network OR health\* OR care\*)`. By integrating the NOT-string, in our search term, we were able to exclude studies, which were not relevant to writing, as feedback, is a broad term, which is also used in computer science, medicine or care work. We nevertheless screened, whether we would erroneously exclude relevant studies by conducting a pilot search, which was not the case in our study. We considered the timespan from January 2003 until December 2020, because around 2003 several tools were either established or got a major upgrade (e.g., MyAccess was established in 2003 and Summary Street was established between 2002 and 2004; Wade-Stein & Kintsch, 2004). To retrieve empirical studies of potential relevance, we performed searches of the following databases: Education Resources Information Center (ERIC; via EBSCOhost), PsychInfo, Web of Science, Google Scholar (first 100 results), PsychArchives, Psycharxiv, and SocArcXiv. The searches yielded in 2,459 hits. We carefully followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses’ (PRISMA) framework of 2020 (Moher et al., 2009; Page et al., 2021; see PRISMA Flow Diagram, see Figure 1). First, the duplicates of the search results were removed. Second, the remaining 2,299 articles were examined by two independent raters to identify studies for inclusion in the systematic review and the meta-analyses. Using the pre-registered inclusion and exclusion criteria (according to the PICOS framework; Fineout-Overholt & Johnston, 2005; Schardt et al., 2007), the articles were first screened by titles and abstracts and then by full texts by two independent raters. We included studies that implemented a feedback tool providing automatically system-generated formative feedback on students’ writing quality to be able to exclude a person's opinion, judgement or expression as a potential influencing factor. Furthermore, the tool should provide feedback on first-language writing drafts to avoid fostering not writing quality, but rather language learning, for example. For the meta-analysis, we only considered studies that followed either a pretest-posttest design or the more rigorous pretest-posttest-control group design to obtain effect sizes and to be able to predict the effectiveness of feedback. We included a broad set of manuscripts to counteract potential effects of publication bias, as long as they were based on empirical research. Articles were excluded when the feedback described in the study was teacher or peer-mediated feedback or when the study was a validation study focusing on the technical functionality of the tool instead of students’ learning outcome. All studies were double coded by a scientific collaborator and the first author of this study (κ = 0.94). After both raters had independently screened all records, the remaining 28 conflicts were resolved through discussion.

Finally, we could derive 24 independent effect sizes for pretest and posttest from 13 of the included articles. Eight of these articles considered a control group. In total, we derived 15 independent effect sizes for studies with a pretest-posttest-control group design (which were also part of the meta-analysis with pretest-posttest design including 24 effects). Thus, we performed a meta-analysis with 24 effect sizes of studies with a pretest-posttest design, and a sub-meta-analysis with the 15 effect sizes of studies with a pretest-posttest-control group design. We deliberately decided against pooling within- and between-effect sizes, as they result from different metrics (measure of change versus measure of group comparisons), and thus, pooling would likely result in distorted findings (see also Morris & DeShon, 2002).

**Figure 2**

*Procedure According to the PRISMA Framework (Inspired by Page et al., 2021)*

## Coding

Three independent raters (κ = 0.67) coded the studies using 33 items collecting information about each study: reference information (e.g., authors, title), sample information (e.g., students’ age, educational level), study conduction (e.g., study setting), feedback information and materials (e.g., feedback tool, text genre). For the meta-analyses the final coding scheme consisted of 19 variables (i.e., authors, scale of the writing quality variable, mean and standard deviation of writing competence in the pretest and posttest for feedback and control condition).

## Research Design and Outcome Measures

In the included studies, students wrote a draft (pretest; e.g., an argumentative essay about a given topic) and after receiving feedback on their writing quality (or no feedback in the control condition) they revised the draft (posttest). Therefore, we applied students’ writing quality in the pretest as independent variable and students’ writing quality in the posttest as dependent variable.

## Moderator Variables

We assessed the impact of the following potential boundary conditions on the effectiveness of computer-based feedback in supporting students’ writing quality during the revision processes: specificity of feedback (generic vs. specific), representation of feedback (e.g., numeric vs. graphical), level of text quality on which feedback was provided (lower vs. higher level), and students’ prior knowledge (see Table 1 for a detailed explanation of the coding of the moderator variables). All moderator variables were coded by three independent raters (κ = 0.73, substantial according to Fleiss’ Kappa). We conducted moderator analyses using the multilevel random effects model.

\*\*\* Insert Table 1 here \*\*\*

## Computing Effect Sizes and Statistical Analyses

### Not Reported Data and Sensitivity Analysis

A common challenge in meta-analyses is dealing with unavailable statistical parameters from the primary studies (Lipsey & Wilson, 2001). Some studies did not provide the minimally required statistical information to conduct a meta-analysis or provided an imprecise operationalization of writing quality. Accordingly, we had to exclude several studies, as we were not able to gather the missing information. To this end, *n* = 24 (with pretest-posttest design and *n* = 15 with pretest-posttest-control group design) of the 48 identified studies could be included in the meta-analysis, as they provided sufficient detail. We used the means, standard deviations, sample sizes, and pretest-posttest correlations to calculate effect sizes (standardized mean change) for the samples of primary studies via the R-package metafor (Viechtbauer, 2010). However, we did not find the pretest-posttest correlation reported in any of the articles. In these cases (single missing statistical information), it is typically suggested to run a sensitivity analysis (Cooper et al., 2019). Sensitivity analyses use a range of plausible values for the missing statistical value, which are then used to continue the data analysis and investigate the variability of results (McKenzie et al., 2022). One of the studies with control group design provided open data (Burkhart et al., 2020). Based on this study we were able to calculate its pretest-posttest correlations (intervention group: *r* = .72; control group: *r* = .86), which we used as an anchor for the sensitivity analysis. We included a broad range of 26 (conservative) plausible values in the sensitivity analyses (cf. Cooper et al., 2019) with correlations from *r* = .50 to .75 for the intervention group samples and correlations from *r* = .64 to .89 for the control group samples, using steps of .1 respectively.

In the results section, we report the mean meta-analytic effect sizes and the according *p*-values. For a better overview and since the variations do not change the result, the ranges of the effect sizes and *p*-values based on the sensitivity analyses are not mentioned in the continuous text, but can be found in the according tables (Tables 3–6).

### Average Effect

For our meta-analysis, we calculated random effect models to account for the heterogeneous design of the included studies. As authors provided multiple comparisons in one paper, we also included the paper-ID as a cluster variable to account for statistical dependencies between effect sizes. We conducted an ANOVA comparing the multilevel random effect model (clustered regarding paper-ID) and the standard model (without cluster variable) to test whether nesting within studies is necessary. The results showed a significant (*p* < .01) preference of the multilevel random effect model over the standard model that indicated that there are substantial differences between papers. Thus, we continued with the random effect model considering paper as cluster variable. As we conducted a meta-analysis on studies with repeated measures (pretest and posttest), we used Becker's (1988) standardized mean change as appropriate effect size measures with qualifying values of *g* = .2, .5, .8 as small, medium, and large effects.

### Publication Bias

To check for possible publication bias, first, we generated a funnel plot (Borenstein et al., 2009; Sterne et al., 2000) in conjunction with trim-and-fill-analyses which proved to be an effective estimation analysis for determining missing studies (Duval & Tweedie, 2000a, 2000b; Pham et al., 2001). In addition, we used Egger’s et al. (1997) regression test to statistically detect the asymmetry of the plot.

### Used Data Management Tools

Data was gathered and organized using Rayaan (<https://www.rayyan.ai/>), a systematic review web application and online tool helping researchers to organize and manage collaboratively the data of systematic reviews and meta-analyses (Ouzzani et al., 2016). Descriptive analyses were carried out using Microsoft Excel. To prepare for statistical analysis, we used SPSS, Version 27. To calculate effect sizes and to conduct the meta-analysis we used the metafor-R-package (Viechtbauer, 2010), which was developed to perform meta-analyses with the R software.

# Results

## Sample

Thirteen articles yielding *k* = 24 pretest-posttest comparisons were included in the meta-analysis (see Table 2). The studies were published between 2013 and 2020 as journal articles (77%), dissertations (15%), or conference papers (8%). Most of the studies were conducted in the United States of America (59%), all others in Germany. There were almost as many laboratory studies (46%) as field studies (55%). A total of *N* = 2,380 students participated in the studies with a mean sample size of *n* = 99 (ranging from *n* = 20 to *n* = 542 participants per study). Most of the participants were university students (64%), followed by students in high school (23%), middle school (9%), and a combined sample of middle and high school students (5%). The text genre in which the students were asked to write their drafts was, in most cases, argumentative/persuasive essay writing (46%), followed by writing explanations (41%), scientific argumentations (9%), and summaries (5%).

Eight of these 24 articles yielded *k* = 15 pairwise comparisons of the writing quality of an experimental group that received feedback with a control group that did not receive feedback, with which we conducted a sub-meta-analysis. The experimental group comprised *n* = 921 participants (ranging from *n* = 13 to *n* = 287 among studies; mean sample size of *n* = 61), the control group comprised *n* = 908 participants (ranging from *n* = 11 to *n* = 270 among studies; mean sample size of *n* = 61), resulting in a total sample size of *N* = 1,829 students included in the sub-meta-analysis.

In Table 2, information about all included studies in both meta-analyses are shown in detail.

\*\*\*Insert Table 2 here\*\*\*

## Publication Bias

As might be expected with large heterogeneity between studies, publication bias revealed significance. The funnel plot of the pretest-posttest meta-analysis shows that the effect sizes are not systematically distributed around the average effect, but there was a gap at the bottom (particularly on the left side) of the plot. The gap indicates that small (nonsignificant) effect sizes of small studies are missing, suggesting that there may be evidence of a possible publication bias. Egger’s regression test (Egger et al., 1997) revealed an asymmetry of the funnel plot, z = 4.1485, *p* < .01, *b* = **−**.1501. However, the trim-and-fill analysis did not include any study but adjusted the center of symmetry. Accordingly, there was a change of the effect size due to the trim-and-fill method resulting in a medium to large significant effect of *g* = .68, *p* < .01, CI 95% [.3787, .9887] (Egger et al., 1997; Sterne et al., 2000).

For the pretest-posttest-control group meta-analysis, there was a different pattern: The funnel plot was relatively symmetrical with a gap at the bottom (right side) of the plot, suggesting that large effect sizes of small studies are missing. However, Egger’s regression test (Egger et al., 1997) revealed an asymmetry of the funnel plot, z = **−**2.3505, *p* = .019, *b* = .7859. Performing the trim-and-fill analysis, two studies were included on the right side and there was a slightly change of the effect, *g* = .37, *p* = .004, CI 95% [.1185, .6294].

In both funnel plots, most studies are shown in the upper range, whereas small studies (with small effects in the pretest-posttest meta-analysis and large effects in the pretest-posttest-control group meta-analysis) are rather absent. This suggests that a so-called small-study effect is present, where the results of smaller studies differ consistently from the larger studies (Schwarzer et al., 2015; Sterne et al., 2000), resulting in an asymmetry. In conclusion, the analyses suggest that the few studies which could be integrated in the present meta-analyses might be report biased. However, all these analyses on publication bias should be treated with caution due to the low number of included studies (Jackson & Turner, 2017). In addition, it should be noted that the asymmetries did probably not result from publication bias but might have other reasons (Egger et al., 1997; Schwarzer et al., 2015). Other possibilities that occur asymmetry of the funnel plot cloud be unbalanced existence of the characteristics in the population (i.e., the population may not be evenly distributed and therefore we should not use an evenly distributed population as the basis for the comparison), small studies showed different (often larger) effects than studies with a larger sample size (small-study effect; Sterne et al., 2000), artefactual correlations of statistical estimates with their standard errors, or only studies with large or significant treatment results were reported (selective/outcome reporting bias; Chan & Altman, 2005).

## Overall Effect Size for Writing Quality

The meta-analysis including 24 independent studies with pretest-posttest design (*N* = 2,380 participants) revealed a standardized mean change of *g* = .53 with *p* = .013, indicating that students improved their writing quality significantly (medium effect) from pretest to posttest. Table 3 shows an overview of the study-wise effect sizes[[1]](#footnote-2) and confidence intervals.

\*\*\*Insert Table 3 here\*\*\*

For the pretest-posttest-control group meta-analysis, we included 15 of the studies (involving 1,829 participants) of the meta-analysis for the pretest-posttest designs only which additionally included a control group. In contrast to the first meta-analysis about the studies with pretest-posttest-design only, under consideration of a control group, we found smaller effects which were not significant: Students in the feedback condition did not improve their writing quality more from pretest to posttest compared to students in the control condition, *g* = .27, *p* = .084 . Table 4 shows an overview of the study-wise effect sizes[[2]](#footnote-3) and confidence intervals.

\*\*\*Insert Table 4 here\*\*\*

## Heterogeneity Analyses

There was statistically observable heterogeneity in the set of effect sizes in both meta-analyses, *Q*(23) = 473.68, *p* < .001, *I*² = 98% (ranging from 97% to 98%) for the meta-analysis for the pretest-posttest design including 24 studies, and *Q*(14) = 117.38, *p* < .001, *I*² = 82% (ranging from 77% to 87%) for the meta-analysis with 15 studies with pretest-posttest-control group design. The *I*² statistic indicated that on average 98% in the pretest-posttest meta-analysis and 82% in the meta-analysis for the pretest-posttest-control group designs of the detected variation could be related to true variation among studies, which refers to high heterogeneity (Higgins et al., 2003). To explore possible reasons of this heterogeneity, we conducted several separate moderator analyses using the multi-level random-effects model.

## Moderator Analyses

For the meta-analysis about 24 effect sizes from studies with pretest-posttest design, several moderators approached significance. Regarding feedback representation, the moderation analyses showed that graphical and highlighting feedback improved students writing quality positively whereas text-based feedback decreased the writing quality (see Table 5). Numeric feedback had no impact on students’ writing quality. Regarding text level, the results indicated that students who received low-level feedback only improved their writing significantly, whereas when students received both low- and high-level feedback their writing quality decreased. High-level feedback only made no difference. Regarding feedback specificity, there was a medium-to-large positive average effect on students’ writing quality indicating that more specific feedback led to larger improvements in student writing quality than generic (non-specific) feedback.

\*\*\*Insert Table 5 here\*\*\*

For the meta-analysis about 15 effect sizes from pretest-posttest-control group design, three moderators revealed significance. Regarding text level, we found that similar to the findings of the pretest-posttest meta-analysis, that only low-level feedback led to higher student writing quality, whereas high-level feedback only or feedback on both lower and higher-level decreased students’ writing quality (see Table 6). Regarding feedback specificity, there was a medium positive average on students’ writing quality suggesting that specific feedback resulted in higher writing quality compared to generic feedback. Regarding prior knowledge, there was a very small negative average effect on students’ writing quality indicating that students with higher prior knowledge profited less from the feedback compared to students with lower prior knowledge.

The effects for the other moderator variables were not significant (see Table 6).

\*\*\*Insert Table 6 here\*\*\*

# Discussion

## Summary of Evidence

The main aim of the present meta-analysis was to investigate the effectiveness of system-generated feedback (compared to no feedback) to support students’ writing quality, and to investigate potential feedback-related and student-related boundary conditions of the effectiveness of computer-based feedback. We found a medium effect of studies with a pretest-posttest design. However, the effect was no longer significant compared to a control group (studies with pretest-posttest-control group design). Thus, overall, we conclude that computer-based feedback is not effective per se.

The higher effects of the pretest-posttest designs can be assumed, as simple pretest-posttest designs rather measure change instead of the effect of an intervention, and also may be prone to other confounds, such as maturation effects (see also Morris, 2008). The effect sizes are in concordance with the meta-analytical findings by Graham et al. (2015), who also documented small to medium effects of computer-based studies, based on a restricted sample of four studies. Additionally, we also obtained a high heterogeneity among the included studies (see also Ngo et al., 2022; Nunes et al., 2021, Strobl et al., 2019, for comparable findings). Part of the variance could be explained by system-related factors: As a consistent pattern across both meta-analyses, we identified the level of text quality and the specificity of feedback as moderators. Low-level feedback increased writing quality more than high-level feedback or feedback that provided both low- and high-level feedback. This finding is surprising, as previous studies suggested that particularly high-level feedback should contribute to writing quality (Crossley & McNamara, 2016; McNamara at al., 2013; Patchan et al., 2016; Strobl et al., 2019;). One explanation could be that the implementation of high-level feedback is presuppositional and requires, for instance, numerical or graphical understanding, high levels of self-regulation and therefore additional assistance during the revision process (Lachner et al., 2017b; McNamara et al., 2015). Therefore, the implementation of low-level feedback could have been less difficult (Berninger & Swanson, 1994; Chanquoy, 2001; Chenoweth & Hayes, 2001; Hacker, 1994; Hayes & Olinghouse, 2015) which could have resulted in a spill-over in high-level implementations and contribute to writing quality (see also Lachner et al., 2017b, for related discussions). However, this interpretation is highly speculative, and requires additional studies, which directly test this hypothesis.

Another interesting finding was that the specificity moderated the effectiveness of computer-based feedback as specific feedback was more effective than general feedback. This finding is consistent with previous research (e.g., Kuklick & Linder, 2023; Lachner et al., 2017a; Mertens et al., 2022; Shute, 2008), and corroborates the need for specific information to implement distinct revision activities to improve writing quality. What constitutes specificity in computer-based feedback, however, goes beyond the scope of this study, as our moderation analyses only allowed to approach the quality of feedback in a coarse-grained manner. Given that our moderation analyses resulted from cross-study comparisons, we want to clarify that these moderation analyses cannot be interpreted in a causal manner.

Regarding the student-related factors, we investigated students’ prior knowledge as potential moderator. In line with previous findings (Fyfe & Rittle-Johnson, 2016; Mertens et al., 2022; Ngo et al., 2022), we found that students with high prior knowledge profited less from the feedback than students with lower prior knowledge. A potential explanation could be the expertise-reversal in the effectiveness of feedback depending on students’ level of knowledge (Kalyuga, 2007; Kalyuga et al., 2003; Kalyuga & Renkl, 2010). That means that feedback that helps less experienced students may inhibit students with higher prior knowledge improving their writing because they retrieve and process redundant information they already know and so impairs their available cognitive capacity. Additional research is needed to investigate how feedback could be implemented to support students with different levels of prior knowledge effectively.

## What are the Theoretical and Practical Contributions of our Meta-Analysis?

Our main findings showed that it is not only about the feedback that matters, but particularly about its boundary conditions (feedback-related as well as student-related). The extent to which computer-based feedback is embedded remains a blind spot. Orchestration could not be investigated in the meta-analysis due to lack of information or transparency in the primary studies, highlighting the need to explore this further in future research.

The small effect sizes of computer-based feedback document that, compared to person-generated feedback by instructors, tutors or peers, computer-based feedback still considerably fell short of expectations, given that computer-based feedback requires distinct, long-term, multi-professional resources during development. In addition, in ill-defined domains such as writing, the effects can be assumed to be smaller, and more fragile than in other, more well-defined domains, such as STEM (Sibley et al., 2023). Large language models, such as GPT or BERT (Scheurer et al., 2022; Tsai et al., 2020; Wu et al., 2022) as well as technical advancements could help provide more sophisticated assessments and feedback approaches, which could additionally contribute to writing quality. More research, however, is necessary to investigate whether and how large language models can be implemented during writing. Another interesting finding for moving the field forward was the restricted number of empirical studies which tapped into the effectiveness of computer-based feedback to foster expository writing (see also Graham et al., 2015). Therefore, we see the need in addition to the development of more advanced computer-based feedback approaches to additionally investigate the effectiveness of the developed technologies with experimental methods to test for causality. These empirical studies would help to gain a better understanding of the conditions of computer-based feedback, and generalize our findings. Last, our findings also have implications for educational practice, as they show that computer-based feedback can be used as a supplement for teachers to foster writing under consideration of the boundary conditions. It can be seen as a general guideline to provide feedback specifically (Mertens et al., 2022; Shute, 2008). Low-level feedback is easy to implement and improves writing quality for less as well as more experienced writers. However, as previous research showed that feedback on higher-level has a larger impact on text quality (Crossley & McNamara, 2016; McNamara et al., 2013; Strobl et al., 2019), we suggest to not only provide low level feedback, but also provide specific high-level feedback, for example by offering additional support to enable students to understand, structure, and process the feedback (Mertens et al., 2022). Additionally, adaptive learning environments offer potential to provide feedback depending on students’ prior knowledge (Kalyuga & Renkl, 2010).

However, there is a lack of field-oriented studies that explore, among other questions, whether it makes a difference whether computer-based feedback is more effective in the classroom or in individual learning sessions. In addition, we see the need to further investigate how computer-based feedback should be orchestrated.

## Limitations

A limitation of this meta-analysis is that we only investigated three system-related and one student-related influencing factor. However, the literature and previous research suggests that there are many more influencing factors (e.g., according to the writing model by Hayes, 2012: environment, motivation, schemes; see also Fleckenstein, Liebenow, & Meyer, 2023; Kluger & DeNisi, 1996; Narciss, 2008; Panadero & Lipnevich, 2022) that should be investigated. In our meta-analysis, first indications emerged that, for instance, cognitive load or mental effort might play a role in the context of more complex feedback or feedback at both lower and higher text levels and also in the context of nonspecific feedback.

Another limitation is that we focused on expository writing in first language and we cannot generalize them on other text genres or foreign language learning. However, previous research showed that feedback tools aiming to foster second language learning provided feedback targeted rather hierarchically lower level of text quality (e.g., spelling and grammar, length, word choice), whereas other feedback tools focused more on content and structure of students’ texts (e.g., overall organization of ideas, cohesion; Strobl et al., 2019). The literature shows also indicators that there are genre-related differences regarding the feedback tools. For example, feedback on academic writing focus more on short and concise sentences and a clear structure of argumentation, whereas feedback on summaries focus more on the appropriateness and completeness of the content, or feedback tools aiming to foster the writing process focus more on self-regulation skills or the use of writing strategies (Nesi & Gardner, 2012; Strobl et al., 2019).

As a final limitation, we have to mention that the sensitivity analysis showed lack of robustness of some effects, and they have to be interpreted with caution. Future research is needed to further increase the validity of our results. However, this requires more open science and transparency of the primary studies. As a first step in this direction, we set ourselves high methodological standards.

## Conclusion

The goal of this meta-analysis was to investigate the effectiveness of computer-based feedback to support students in expository writing with aggregate data. Furthermore, we aimed at identifying potential boundary conditions to complement the existing research and gain new insights for practitioners as well as for researchers into computer-based formative feedback to improve expository writing automatically. Overall, the meta-analysis presented in this paper provides important evidence that specific computer-based feedback particularly on lower text level improved students’ writing quality, especially for less experienced writers.

# References

Articles included in the meta-analyses are marked with an asterisk.

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. Learning and Instruction, 16(3), 183–198. <https://doi.org/10.1016/j.learninstruc.2006.03.001>

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.). *Handbook of writing research* (2nd ed., pp. 316–329). Guilford.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*(2), 257–278. <https://doi.org/10.1111/j.2044-8317.1988.tb00901.x>

Becker-Mrotzek, M., Grabowski, J., Jost, J., Knopp, M., & Linnemann, M. (2014). Adressatenorientierung und Kohärenzherstellung im Text [Audience focus and generating coherence in the text]. *Didaktik Deutsch, 19*(37), 21–43.<https://doi.org/10.25656/01:17151>

Bereiter, C., & Scardamalia, M. (1987). The psychology of written composition. Erlbaum.

Berninger, V. W., & Swanson, H. L. (1994). Modification of the Hayes and Flower model to explain beginning and developing writing. In E. Butterfield (Ed.), *Advances in Cognition and Educational Practice* (Vol. 2, pp. 57–82). JAI Press.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Bromme, R., Jucks, R., & Rambow, R. (2004). Experten-Laien-Kommunikation im Wissensmanagement [Communication of experts and novices in knowledge management]. In G. Reinmann & H. Mandl (Eds.). Der Mensch im Wissensmanagement: Psychologische Konzepte zum besseren Verständnis und Umgang mit Wissen [The human being in knowledge management: Psychological concepts for a better understanding and handling of knowledge] (pp. 176–188). Hogrefe.

\*Burkhart, C., Lachner, A., & Nückles, M. (2020). Using spatial contiguity and signaling to optimize visual feedback on students’ written explanations. *Journal of Educational Psychology*. [https://doi.org/10.1037/edu0000607](https://psycnet.apa.org/doi/10.1037/edu0000607)

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

Chan, A. W., & Altman, D. G. (2005). Epidemiology and reporting of randomised trials published in PubMed Journals. *The Lancet*, *365*, 1159–1162.   
<http://dx.doi.org/10.1016/S0140-6736(05)71879-1>

Chanquoy, L. (2001). How to make it easier for children to revise their writing? A study of text revision from 3rd to 5th grades. *British Journal of Educational* *Psychology*, *71*, 15–41. <https://doi.org/10.1348/000709901158370>

Chanquoy, L. (2009). Revision processes. In R. Beard, D. Myhill, J. Riley, & M. Nystrand (Eds.), *The SAGE handbook of writing development* (pp. 80–97). SAGE.

Chenoweth, N., & Hayes, J. R. (2001). Fluency in writing. Generating text in L1 and L2. Written Communication, *18*, 80–98. <https://doi.org/10.1177/0741088301018001004>

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 328–338. <https://doi.org/10.1016/j.learninstruc.2009.08.006>

Concha, S., & Paratore, J. R. (2010). Local coherence in persuasive writing: An exploration of Chilean students’ metalinguistic knowledge, writing process, and writing products. *Written Communication*, *28*(1), 34–69.  [https://doi.org/10.1177/0741088310383383](%20https://doi.org/10.1177/0741088310383383)

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation. <https://doi.org/10.7758/9781610448864>

Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How elaboration and cohesion can increase writing quality. *The Journal of Writing Research*, *7*(3), 351–370. https://doi.org/10.17239/jowr-2016.07.03.02

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1). <https://ejournals.bc.edu/index.php/jtla/article/view/1640>

Duval, S. J., & Tweedie, R. L. (2000a). A nonparametric "trim and fill" method of  
accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*(449), 89–98. <https://doi.org/10.1080/01621459.2000.10473905>

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

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal,* *315*(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>

Fineout-Overholt, E., & Johnston, L. (2005). Teaching EBP: asking searchable, answerable clinical questions. *Worldviews On Evidence-Based Nursing*, *2*(3), 157–160. <https://doi.org/10.1111/j.1741-6787.2005.00032.x>

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feednack and writing: A multi-level meta-analysis of effects on students’ performance. *Frontiers in Artificial Intelligence*, *6*, 1162454. <https://doi.org/10.3389/frai.2023.1162454>

Fleckenstein, J., Reble, R., Meyer, J., Jansen, T., Liebenow, L. W., Möller, J., & Köller, O. (2023). Digitale Schreibförderung im Bildungskontext: Ein systematisches Review [Digital writing support in educational contexts: A systematic review]. In K. Scheiter & I. Gogolin (Eds.), *Bildung für eine digitale Zukunft* [Education for a digital future] (Edition ZfE, Vol. 15, pp. 3–25). Springer VS. <https://doi.org/10.1007/978-3-658-37895-0_1>

Flower, L., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. Gregg, & E. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31–50). Erlbaum.

\*Frost, K. L. (2008). The effects of automated essay scoring as a high school classroom intervention [Doctoral dissertation, University of Nevada, Las Vegas]. *UNLV Retrospective Theses & Dissertations*, 2839. <http://dx.doi.org/10.25669/37p4-kdh7>

Fyfe, E. R., & Rittle-Johnson, B. (2016). Feedback both helps and hinders learning: The causal role of prior knowledge. *Journal of Educational Psychology*, *108*, 82–97. <https://doi.org/10.1037/edu0000053>

Geiser, S., & Studley, R. (2001). *UC and SAT: Predictive validity and differential impact of the SATI and SAT II at the University of California*. University of California. <https://doi.org/10.4324/9780203463932_uc_and_the_sat>

Goodman, J., Wood, R. E., & Hendrickx, M. (2004). Feedback specificity, exploration, and learning. *Journal of Applied Psychology*, *89*, 248–262. [https://doi.org/10.1037/0021-9010.89.2.248](https://psycnet.apa.org/doi/10.1037/0021-9010.89.2.248)

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education,* *43*(1), 277–303. <https://doi.org/10.3102/0091732X18821125>

Graham, S., Harris, K. R., & Hebert, M. (2011). *Informing writing: The beneﬁts of formative assessment.* Alliance for Excellence in Education.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, *115*(4), 523–547. <https://doi.org/10.1086/681947>

Hacker, D. J. (1994). Comprehension monitoring as a writing process. In E. C. Butterfield (Ed.), *Children’s writing: Toward a process theory of the development of skilled writing*. JAI Press.

Harris, K. R., Santangelo, T., & Graham, S. (2008). Self-regulated strategy development in writing: Going beyond NLEs to a more balanced approach. *Instructional Science*, *36*, 395–408. <https://doi.org/10.1007/s11251-008-9062-9>

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

Hayes, J. R. (2012). Modeling and Remodeling Writing. *Written Communication*, *29*(3), 369–388. <https://doi.org/10.1177/0741088312451260>

Hayes, J. R., & Flower, L. S. (1986). Writing research and the writer. *American Psychologist*, *41*(10), 1106–1113. <https://doi.org/10.1037/0003-066x.41.10.1106>

Hayes, J. R., & Olinghouse, N. G. (2015). Can cognitive writing models inform the design of the common core state standards? *The Elementary School Journal*, *115*(4), 480–497. <https://doi.org/10.1086/681909>

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed.)*, *327*(7414), 557–560. <https://doi.org/10.1136/bmj.327.7414.557>

Jackson, D., & Turner, R. (2017). Power analysis for random-effects meta-analysis. *Research Synthesis Methods*, *8*(3), 290–302. <https://doi.org/10.1002/jrsm.1240>

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review,* *19*, 509–539. <https://doi.org/10.1007/s10648-007-9054-3>

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect, *Educational Psychologist*, *38*(1), 23–31. <https://doi.org/10.1207/S15326985EP3801_4>

Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science,* *38*, 209–215. <https://doi.org/10.1007/s11251-009-9102-0>

Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review*, *14*(2), 237–242. <https://doi.org/10.3758/bf03194058>

Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist, 44*(4), 250–266. <https://doi.org/10.1080/00461520903213600>

\*Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write?, *Journal of Educational Computing Research*, *42*(2), 173–196. <https://doi.org/10.2190/EC.42.2.c>

Kieft, M., Rijlaarsdam, G., & van den Bergh, H. (2008). An aptitude–treatment interaction approach to writing-to-learn. *Learning and Instruction*, *18*(4), 379–390. <https://doi.org/10.1016/j.learninstruc.2007.07.004>

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

\*Lachner, A., Burkhart, C., & Nückles, M. (2017a). Formative computer-based feedback in the university classroom: Specific concept maps scaffold students’ writing. *Computers in Human Behavior, 72*(4), 459–469. <http://dx.doi.org/10.1016/j.chb.2017.03.008>

\*Lachner, A., Burkhart, C., & Nückles, M. (2017b). Mind the gap! Automated concept map feedback supports students in writing cohesive explanations. *Journal of Experimental Psychology: Applied, 23*(1), 29–46. <http://dx.doi.org/10.1037/xap0000111>

\*Lachner, A., & Neuburg, C. (2019). Learning by writing explanations: computer-based feedback about the explanatory cohesion enhances students’ transfer. *Instructional Science*, *47*(1), 19–37.  [https://doi.org/10.1007/s11251-018-9470-4](%20https://doi.org/10.1007/s11251-018-9470-4)

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. Sage Publications, Inc.

MacArthur, C. A., & Graham, S. (2016). Writing research from a cognitive perspective. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), Handbook of writing research (pp. 24–40). The Guilford Press.

\*McCarthy, K. S., Roscoe, R. D., Likens, A. D., & McNamara, D. S. (2019). Checking it twice: Does adding spelling and grammar checkers improve essay quality in an automated writing tutor? In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education, 20th International Conference, AIED 2019 Chicago, IL, USA, June 25–29, 2019 Proceedings, Part I* (pp. 270–282). Springer. <https://doi.org/10.1007/978-3-030-23204-7>

McKenzie, J. E., Brennan, S. E., Ryan, R. E., Thomson, H. J., & Johnston, R. V. (2022). Analysing data and undertaking meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.). *Cochrane Handbook for Systematic Reviews of Interventions* (Version 6.3). Cochrane.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*, 499–515. <https://doi.org/10.3758/s13428-012-0258-1>

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59. <https://doi.org/10.1016/j.asw.2014.09.002>

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), 1743–1772. <https://doi.org/10.1037/edu0000764>

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Med*, *6*(7). <https://doi.org/10.1371/journal.pmed.1000097>

Moore, N. S., & MacArthur, C. A. (2016). Student use of automated essay evaluation technology during revision. *Journal of Writing Research*, *8*. <http://dx.doi.org/10.17239/jowr-2016.08.01.05>

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364–386. [https://doi.org/10.1177/1094428106291059](https://psycnet.apa.org/doi/10.1177/1094428106291059)

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125. <https://doi.org/10.1037//1082-989X.7.1.105>

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–144). Lawrence Erlbaum.

Nesi, H., & Gardner, S. (2012). Families of genres of assessed writing. In H. Nesi & S. Gardner (Eds). *Genres across the disciplines: Student writing in higher education* (pp. 21–56). Cambridge University Press. <https://doi.org/10.1017/9781009030199>

Ngo, T. T-N., Chen, H. H.-J., & Lai, K. K.-W. (2022). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments*, *1*(1), 1–18. <https://doi.org/10.1080/10494820.2022.2096642>

Nihalani, P. K., Mayrath, M., & Robinson, D. H. (2011). When feedback harms and collaboration helps in computer simulation environments: An expertise reversal effect. *Journal of Educational Psychology, 103*(4), 776–785. <https://doi.org/10.1037/a0025276>

Nückles, M., Winter, A., Wittwer, J., Herbert, M., & Hübner, S. (2006). How do experts adapt their explanations to a layperson’s knowledge in asynchronous communication? An experimental study. *User Modeling and User-Adapted Interaction, 16*(2), 87–127. <https://doi.org/10.1007/s11257-006-9000-y>

Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2021). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, *38*(2), 599–620. <https://doi.org/10.1111/jcal.12635>

O’Rourke, S. T., Calvo, R. A., & McNamara, D. S. (2011). Visualizing topic flow in students’ essays. *Journal of Educational Technology &* *Society, 14,* 4–15.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016)*.* Rayyan – A web and mobile app for systematic reviews. *Systematic Reviews, 5*, 210. <https://doi.org/10.1186/s13643-016-0384-4>

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

\*Palermo, C. J. (2017). A framework for deliberate practice: Self-regulated strategy development and an automated writing evaluation program [Doctoral dissertation, North Carolina State University]. <http://pqdtopen.proquest.com/#viewpdf?dispub=10610664>

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, *35*, 100416. <https://doi.org/10.1016/j.edurev.2021.100416>

Patchan, M., Schunn, C., & Correnti, R. (2016). The nature of feedback: how peer feedback features affect students’ implementation rate and quality of revisions. *Journal of* *Educational Psychology*, *108*(8), 1098–1120. [https://doi.org/10.1037/edu0000103](https://psycnet.apa.org/doi/10.1037/edu0000103)

Pham, B., Platt, R., McAuley, L., Klassen, T. P., & Moher, D. (2001). Is there a  
"best" way to detect and minimize publication bias? An empirical evaluation. *Evaluation & the Health Professions*, *24*(2), 109–125. <https://doi.org/10.1177/016327870102400202>

Philipp, M. (2014). *Selbstreguliertes Schreiben. Schreibstrategien erfolgreich vermitteln [Self-regulated writing. Teaching writing strategies successfully]*.Beltz.

Philipp, M. (2017). Wirksame Schreibförderung – Metaanalytische Befunde [Effective writing support - Meta-analytic findings]. In M. Becker-Mrotzek, J., Grabowski, & T. Steinhoff, (Eds.), *Forschungshandbuch empirische Schreibdidaktik (*pp. 187–202). Waxmann.

Pirnay-Dummer, P., & Ifenthaler, D. (2011). Reading guided by automated graphical representations: How model-based text visualizations facilitate learning in reading comprehension tasks. Instructional Science, 39(6), 901–919. [https://doi.org/10.1007/s11251-010-9153-2](https://psycnet.apa.org/doi/10.1007/s11251-010-9153-2)

Powell, P. (2009). Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication*, *60*, 664–682.

Richter, J., Scheiter, K., & Eitel, A. (2016). Signaling text-picture relations in multimedia learning: A comprehensive meta-analysis. *Educational Research Review*, *17*, 19–36. <https://doi.org/10.1016/j.edurev.2015.12.003>

Roscoe, R. D., Jacovina, M. E., Allen, L. K., Johnson, A. C., & McNamara, D. S. (2016). Toward revision-sensitive feedback in automated writing evaluation. *Educational Data Mining*, 628–629.

Roscoe, R. D., & McNamara, D. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, *105*, 1010–1025. <https://doi.org/10.1037/a0032340>

\*Roscoe, R. D., Snow, E. L., & McNamara, D. S. (2013). Feedback and revising in an intelligent tutoring system for writing strategies. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education*. *Lecture Notes in Computer Science* (Vol. 7926, pp. 259–268). Springer. <https://doi.org/10.1007/978-3-642-39112-5_27>

Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? International Journal of Human-Computer Studies, 45(2), 185–213. [https://doi.org/10.1006/ijhc.1996.0048](https://psycnet.apa.org/doi/10.1006/ijhc.1996.0048)

Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making*, *7*(16). <https://doi.org/10.1186/1472-6947-7-16>

Scheurer, J., Campos, J. A., Chan, J. S., Chen, A., Cho, K., & Perez, E. (2022). Training language models with natural language feedback. arXiv:2204.14146. <https://doi.org/10.48550/arXiv.2204.14146>

Schneider, J., Backfisch, I., & Lachner, A. (2021). Facilitating open science practices for research syntheses: PreregRS guides preregistration. *Research Synthesis Methods*, *13*(2), 284–289. <https://doi.org/10.1002/jrsm.1540>

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R.* Springer. <https://doi.org/10.1007/978-3-319-21416-0>

Sharp, L. A. (2016). Preservice teachers and writing: An analysis of academic performance—RESEARCH. *Kentucky Journal of Excellence in College Teaching and Learning*, *13*(2). <https://encompass.eku.edu/kjectl/vol13/iss1/2>

Shermis, M. D., & Burstein, J. (Eds.). (2003). Automated essay scoring: A cross-disciplinary perspective. Lawrence Erlbaum Associates Publishers.

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. <https://doi.org/10.3102/0034654307313795>

Sibley, L., Lachner, A., Plicht, C., Fabian, A., Backfisch, I., Scheiter, K., & Bohl, T. (2023). *Feasibility of adaptive teaching with technology: Which implementation conditions matter?* [Manuscript submitted for publication]. Institute of Education, University of Tübingen.

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis. *Journal of Clinical Epidemiology*, *53*(11), 1119–1129. <https://doi.org/10.1016/S0895-4356(00)00242-0>

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing, *Assessing Writing*, 19, 51–65. <https://doi.org/10.1016/j.asw.2013.11.007>

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies, *Computers and Education*, *131*, 33–48. <https://doi.org/10.1016/j.compedu.2018.12.005>

Tsai, C.-T., Chen, J.-J., Yang, C.-Y., & Chang, J. S. (2020). LinggleWrite: A Coaching System for Essay Writing. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. <https://doi.org/10.18653/v1/2020.acl-demos.17>

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

Villalon, J., & Calvo, R. A. (2011). Concept maps as cognitive visualizations of writing assignments. *Journal of Educational Technology & Society*, *14*(3), 16–27.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction*, *3*, 333–362. <https://doi.org/10.1207/s1532690xci2203_3>

\*Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., Magooda, A., & Quintana, R. (2020). eRevis(ing): Students’ revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, *44*, 100449. <https://doi.org/10.1016/j.asw.2020.100449>

\*Weston-Sementelli, J. L., Allen, L. K., & McNamara, D. S. (2016). Comprehension and Writing Strategy Training Improves Performance on Content-Specific Source-Based Writing Tasks. *International Journal of Artificial Intelligence in Education*, 1–32. <https://doi.org/10.1007/s40593-016-0127-7>

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metricsof efficacy. *Journal of Educational Computing Research*, *58*, 87–125. <https://doi.org/10.1177/0735633119830764>

Winne, P. H., & Butler, D. L. (1994). Student Cognition in learning from teaching. In T. Husen & T. Postlewaite (Eds.), *International Encyclopedia of Education* (2nd ed., pp. 5738–5745). Pergamon.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research, *Frontiers in Psychology*, *10*(3087). <https://doi.org/10.3389/fpsyg.2019.03087>

Wu, Y., Henriksson, A., Nouri, J., Duneld, M., & Li, X. (2022). Beyond benchmarks: Spotting key topical sentences while improving automated essay scoring performance with topic-aware BERT. *Electronics*, *12*(1), 150. <https://doi.org/10.3390/electronics12010150>

Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, *61*(4), 875-900. <https://doi.org/10.1177/07356331221127300>

\*Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668. <https://doi.org/10.1016/j.compedu.2019.103668>

\*Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students’ scientific argumentation. *International Journal of Science Education,* 1648–1668. <https://doi.org/10.1080/09500693.2017.1347303>

# Tables

**Table 1**

*Moderator Variables*

| Variable | Description | Scales | Prototypical example of Coding |
| --- | --- | --- | --- |
| Text level | Feedback on a lower hierarchical level focuses on surface features of the text (e.g., grammar, spelling, text length, word count). Feedback on a higher hierarchical level is more complex and has a larger influence on writing quality and text comprehension (e.g., organization and structure, style, cohesion). | 1 = Low-level only  2 = High-level only  3 = Both lower and higher level | In the study by Roscoe et al. (2015), feedback was provided on different text properties ranging from lower level aspects, e.g., number of words and sentences, to more sophisticated aspects on higher level, e.g., cohesion. Thus, this study was coded with 3 (*both lower and higher level*). |
| Graphical feedback representation | Feedback is visualized graphical, i.e., in form of a concept map. | 0 = Not graphical  1 = Graphical | Burkhart et al. (2020) provided one group of students feedback in form of a conventional concept map which was coded with 1 as graphical representation. |
| Numeric feedback representation | Feedback is visualized numeric, e.g., as scores, grades, points achieved, percentages, diagrams, or rating scales. | 0 = Not numeric  1 = Numeric | In the study by Kellogg et al. (2010), feedback was provided in form of a holistic score. Thus, this study was coded with 1 (*numeric representation of feedback*). |
| Highlighting representation of feedback | Feedback information is directly embedded within the text through signaling, markings, or symbols. | 0 = No highlighting  1 = Highlighting was used | In the study by Burkhart et al. (2020), in the signaled concept map condition, signaling was used embedded directly in students’ draft and within the concept map. This case was coded with 1 as *highlighting* was used to represent the feedback information. |
| Text-based feedback representation | Feedback is represented as text-based comment, hints, massages, or suggestions. | 0 = Not text-based  1 = Text-based | In the study by Lachner et al. (2017a), in the outline condition, the feedback information was represented only as keywords why is was coded with 1 (*text-based*). |
| Number of representations | Number of representations describes whether only a single representation format was used to provide the feedback or whether multiple representation formats were combined. | 0 = Mono (single) representation  1 = Multiple representations | In the study by Lachner et al. (2017b), feedback regarding local and global cohesion was provided by using a concept map (graphical). Thus, this study was coded with 0 *(mono representation)*. In contrast, in the study by Palermo et al. (2017), feedback was presented in form of scores in bar charts (numeric) combined with written statements (text-based). Thus, this study was coded with 1 (*multiple representations*). |
| Feedback specificity | Feedback specificity describes how detailed and comprehensive feedback is provided. In generic feedback, the writing product is considered as a unit (overall score). In specific feedback procedure, feedback is given on different criteria of the text. | 1 = Generic (only one generic score or one generic feedback method is provided)  2 = Specific (there are more than one feedback representation and possibly additional suggestions or prompts to help implementing the (localized) feedback are provided) | In the study by Wang et al. (2020), students receive a grade on their first and a grade on their second draft as feedback. Thus, this study was coded with 1 (*generic*). In contrast, in the study by Wilson et al. (2020), 6 traits of writing quality were scored on a scale from 1 to 5 and students were provided with feedback on each of the 6 traits, indicating strategies for students to improve their draft regarding the specific trait. Thus, this study was coded with 2 (*specific*). |
| Prior knowledge | Students’ prior knowledge was measured using the pretest measurements. | Because the different studies used different scales to measure writing quality, we transformed students’ pretest values into percentages so that we could measure the moderator variable on a scale from 0% to 100% | In the study by Zhu et al. (2017), they used for students’ pretest measures a scale from 0 to 10. The mean writing quality was 5.20. We translated the values into percentage ( resulted in 52%. |
| Setting |  | 1 = Laboratory  2 = Field study |  |
| Educational level |  | 0 = School (including middle and high school)  1 = University |  |

**Table 2**

*Overview of all Included Studies in the Meta-Analyses*

| Study | Type of Manus-cript | Country | Sample size in Total | Level of Education | Study Design | Sample size per Condition | | Feedback tool | Type of Feedback | Text Level | Feedback Representation | Feedback Specificity | Study Setting | Text Genre |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | *n*  Experi-mental Group | *n*  Control Group |  |  |  |  |  |  |  |
| Burkhart et al., 2020 | journal article | Germany | 100 | university | pre-post-control | 25 | 26 | CohViz | correspondence-enhanced concept-map feedback | higher | Highlighting in the text + concept map | specific | laboratory | explanation |
| journal article | Germany | 100 | university | pre-post-control | 23 | 26 | CohViz | conventional concept-map feedback | higher | Concept map | specific | laboratory | explanation |
| journal article | Germany | 100 | university | pre-post-control | 26 | 26 | CohViz | spatially contiguous feedback | higher | Highlighting in the text + symbols | specific | laboratory | explanation |
| Frost, 2008 | disser-tation | USA | 31 | high school | pre-post-control | 23 | 11 | Criterion | AES feedback + teacher instruction | Lower + higher | numeric | generic | field | summary |
| McCarthy et al., 2019 | journal article | USA | 60 | high school | pre-post | - | - | Writing  Pal | Writing Pal feedback (generic score) | Lower + higher | Numeric + text-based | generic | laboratory | (argumentative / persuasive) essay |
| journal article | USA | 59 | high school | pre-post | - | - | Writing  Pal | writing Pal feedback + spelling and grammar checking (generic score) | Lower + higher | Numeric + text-based | generic | laboratory | (argumentative / persuasive) essay |
| Kellogg et al., 2010 | journal article | USA | 39 | university | pre-post-control | 20 | 19 | Criterion | intermittent feedback | Lower + higher | numeric | generic | laboratory | (argumentative / persuasive) essay |
| journal article | USA | 20 | university | pre-post-control | 20 | 19 | Criterion | continuous feedback | Lower + higher | numeric | generic | laboratory | (argumentative / persuasive) essay |
| Lachner et al., 2017a | journal article | Germany | 96 | university | pre-post | - | - | CohViz | general feedback | higher | numeric | specific | field | explanation |
| journal article | Germany | 77 | university | pre-post | - | - | CohViz | concept map feedback | higher | Concept map | specific | field | explanation |
| journal article | Germany | 78 | university | pre-post | - | - | CohViz | outline feedback | higher | text | specific | field | explanation |
| Lachner et al. (2017b), Study 2 | journal article | Germany | 42 | university | pre-post-control | 21 | 21 | CohViz | concept map feedback | higher | Concept map | specific | laboratory | explanation |
| Lachner et al. (2017b), Study 3 | journal article | Germany | 27 | university | pre-post-control | 13 | 14 | CohViz | concept map feedback | higher | Concept map | specific | laboratory | explanation |
| Lachner & Neuburg, 2019 | journal article | Germany | 61 | university | pre-post-control | 31 | 30 | CohViz | concept map feedback with signaling | higher | Highlighting in the text + concept map | specific | laboratory | explanation |
| Palermo, 2017 | disser-tation | USA | 542 | middle school | pre-post-control | 272 | 270 | NCWrite | feedback + traditional instruction | lower | Numeric + text-based | specific | field | (argumentative / persuasive) essay |
| disser-tation | USA | 557 | middle school | pre-post-control | 287 | 270 | NCWrite | feedback + self-regulated strategy development | lower | Numeric + text-based | specific | field | (argumentative / persuasive) essay |
| Roscoe et al., 2013 | con-ference paper | USA | 65 | high school | pre-post-control | 33 | 32 | Writing  Pal | Writing-Pal condition | Lower + higher | Numeric + text-based | generic | field | (argumentative / persuasive) essay |
| Wang et al., 2020 | journal article | USA | 143 | middle school | pre-post |  |  | eRevise | AWE system eRevise | lower | numeric | generic | field | (argumentative / persuasive) essay |
| Weston-Sementelli et al., 2016 | journal article | USA | 89 | university | pre-post-control | 41 | 48 | Writing  Pal | iSTART | Lower + higher | Numeric + text-based | generic | field | (argumentative / persuasive) essay |
| journal article | USA | 41 | university | pre-post-control | 41 | 48 | Writing  Pal | Writing-Pal | Lower + higher | Numeric + text-based | generic | field | (argumentative / persuasive) essay |
| journal article | USA | 45 | university | pre-post-control | 45 | 48 | Writing  Pal | iSTART + Writing-Pal | Lower + higher | Numeric + text-based | generic | field | (argumentative / persuasive) essay |
| Zhu et al., 2020 | journal article | USA | 131 | high + middle school | pre-post | - | - | c-rater ML | generic | lower | Numeric + text-based + scales | specific | field | scientific argumentation |
| journal article | USA | 203 | high + middle school | pre-post | - | - | c-rater ML | contextualized | higher | Numeric + text-based + scales | specific | field | scientific argumentation |
| Zhu et al., 2017 | journal article | USA | 141 | high school | pre-post | - | - | c-rater ML | automated scoring feedback | higher | Numeric + text-based | specific | field | scientific argumentation |

**Table 3**

*Overview of Effect Sizes and Confidence Intervals of all Included Studies in the Meta-Analysis of Studies with Pretest-Posttest Design (k = 24)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Study | Number of participants *n* | Effect size *g* | *SE* | 95% CI | |
| Lower | Upper |
| Burkhart et al., 2020 | 51 | 1.47 | .270 (.251-.288) | .9395 (.9040-.9766) | 1.9990 (1.9619-2.0345) |
| Burkhart et al., 2020 | 49 | .72 | .209 (.181-.234) | .3085 (.2593-.3619) | 1.1266 (1.0733-1.1759) |
| Burkhart et al., 2020 | 52 | .61 | .189 (.162-.213) | .2358 (.1878-.2882) | .9762 (.9238-1.0242) |
| Frost, 2008 | 34 | −.54 | .197 (.167-.223) | −.9235 (−.9755-−.8665) | −.1530 (−.2101-−.1010) |
| Kellogg et al., 2010 | 39 | 2.69 | .467 (.453-.480) | 1.7725 (1.7465-1.7990) | 3.6030 (3.5765-3.6290) |
| Kellogg et al., 2010 | 39 | 2.95 | .505 (.492-.517) | 1.9590 (1.9349-1.9835) | 3.9376 (3.9131-3.9617) |
| Lachner et al., 2017a | 96 | .13 | .088 (.073-.102) | −.0428 (−.0704-−0.0121) | .3039 (.2732-.3314) |
| Lachner et al., 2017a | 77 | .13 | .099 (.081-.114) | −.0616 (−.0923-−0.0273) | .3256 (.2913-.3563) |
| Lachner et al., 2017a | 78 | .19 | .099 (.082-.114) | −.0031 (−.0335-.0307) | .3840 (.3502-.4144) |
| Lachner et al., 2017b (Study 2) | 42 | 1.43 | .290 (.269-.310) | .8598 (.82044-.9010) | 1.9969 (1.9557-2.0362) |
| Lachner et al., 2017b (Study 3) | 27 | .70 | .276 (.240-.310) | .1609 (.0951-.2323) | 1.2430 (1.1716-1.3088) |
| Lachner & Neuburg, 2019 | 61 | 1.04 | .203 (.183-.223) | .6379 (.6002-.6781) | 1.4354 (1.3952-1.4731) |
| McCarthy et al., 2019 | 60 | .22 | .113 (.094-.131) | .0027 (−0.0318-.0410) | .4461 (.4077-.4806) |
| McCarthy et al., 2019 | 59 | .22 | .114 (.094-.132) | −.0074 (−0.0422-.0313) | .4392 (.4005-.4740) |
| Palermo, 2017 | 542 | .44 | .056 (.047-.064) | .3333 (.3178-.3504) | .5513 (.5342-.5669) |
| Palermo, 2017 | 557 | .97 | .065 (.058-.072) | .8436 (.8309-.8572) | 1.0989 (1.0853-1.1116) |
| Roscoe et al., 2013 | 65 | .14 | .151 (.124-.175) | −.1564 (−.2034-−0.1041) | .4354 (.3831-.4823) |
| Wang et al., 2020 | 143 | .15 | .073 (.060-.084) | .0035 (−0.0190-.0287) | .2880 (.2629-.3105) |
| Weston-Sementelli et al., 2016 | 89 | −.18 | .136 (.112-.157) | −.4507 (−.4926-−.4040) | .0828 (.0361-.1247) |
| Weston-Sementelli et al., 2016 | 89 | .04 | .135 (.111-.156) | −.2261 (−.2685-−.1789) | .3016 (.2543-.3439) |
| Weston-Sementelli et al., 2016 | 93 | .65 | .146 (.126-.164). | .3654 (.3295-.4046) | .9368 (.8977-.9728) |
| Zhu et al., 2020 | 131 | 1.87 | .138 (.131-.145) | 1.6023 (1.5890-.1.6160) | 2.1440 (2.1303-2.1574) |
| Zhu et al., 2020 | 203 | 1.39 | .092 (.085-.010) | 1.2064 (1.1935-1.2199) | 1.5659 (1.5524-1.5788) |
| Zhu et al., 2017 | 141 | .35 | .076 (.063-.087) | .2036 (.1816-.2280) | .4998 (.4754-.5218) |

*Note.**g* is the effect size of Becker’s (1988) standardized mean change from students’ first draft (pretest) to their revised draft (posttest); the effect size did not vary due to different pretest-posttest correlation. *SE* is the standard error of the effect size. 95% CI is a 95% confidence interval with its lower and upper limit. The table shows the mean effect size, its standard error, confidence interval, and the corresponding *p*-value across the 26 meta-analyses based on the mean correlation according to the sensitivity analysis procedure. The range of the respective values of the 26 meta-analyses with the smallest and largest correlation is reported in parentheses. Effect size does not vary due to different pretest-posttest correlation, hence only one effect size value per study is reported.

**Table 4**

*Effect Sizes and Confidence Intervals of all Included Studies in the Meta-Analysis of Studies with Pretest-Posttest-Control Group-Design (k = 15)*

| Study | Number of participants *n* | Effect size *g* | *SE* | 95% CI | |
| --- | --- | --- | --- | --- | --- |
| *LL* mean (min-max) | *UL* mean (min-max) |
| Burkhart et al., 2020 | 51 | .68 | .321 (.289-.350) | .05283 (−.0054-.1148) | 1.3100 (1.2480-1.3683) |
| Burkhart et al., 2020 | 49 | −.07 | .271 (.231-.307) | −.6009 (−.6722-−.5229) | .4604 (.3824-.5316) |
| Burkhart et al., 2020 | 52 | −.18 | .256 (.216-.292) | −.6831 (−.7538-−.6053) | .3195 (.2417-.3901) |
| Frost, 2008 | 34 | −.06 | .281 (.216-.337) | −.6075 (−.7182-−.4802) | .4939 (.3665-.6045) |
| Kellogg et al., 2010 | 39 | −1.07 | .784 (.768-.901) | −2.6118 (−2.6436-−2.5795) | .4629 (.4306-.4948) |
| Kellogg et al., 2010 | 39 | −.81 | .807 (.791-.823) | −2.3964 (−2.4273843-−2.3651) | .7687 (.7373-.7996) |
| Lachner et al., 2017b (Study 2) | 42 | .87 | .337 (.300-.371) | .2120 (.1449-.2838) | 1.5331 (1.4613-1.6002) |
| Lachner et al., 2017b (Study 3) | 27 | .49 | .333 (.273-.386) | −.1651 (−.2693-−.0488) | 1.1394 (1.0231-1.2436) |
| Lachner & Neuburg, 2019 | 61 | .38 | .253 (.219-.284) | −.1192 (−.1804-−.0527) | .8730 (.8066-.9342) |
| Palermo, 2017 | 542 | .64 | .070 (.056-.082) | .5040 (.4794-.5318) | .7774 (.7496-.8020) |
| Palermo, 2017 | 557 | 1.17 | .078 (.065-.089) | 1.0175 (.9958-1.0415) | 1.3217 (1.2978-1.3435) |
| Roscoe et al., 2013 | 65 | −.02 | .194 (.151-.231) | −.4029 (−.4765-−.3186) | .3567 (.2724-.4303) |
| Weston-Sementelli et al., 2016 | 89 | −.07 | .168 (.132-.200) | -.4025 (-.4649-−.3312) | .2559 (.1847-.3184) |
| Weston-Sementelli et al., 2016 | 89 | .15 | .167 (.130-.199) | −.1784 (−.2413-−.1065) | .4753 (.4034-.5381) |
| Weston-Sementelli et al., 2016 | 93 | .76 | .176 (.143-.205) | .4170 (.3600-.4810) | 1.1067 (1.0427-1.1638) |

*Note.**g* is the effect size of Becker’s (1988) standardized mean change from students’ first draft (pretest) to their revised draft (posttest) in comparison with a control group, *SE* is the standard error of the effect size, 95% CI is a 95% confidence interval with its lower and upper limit. The table shows the mean effect size, its standard error, confidence interval, and the corresponding *p*-value across the 26 meta-analyses based on the mean correlation according to the sensitivity analysis procedure. The range of the respective values of the 26 meta-analyses with the smallest and largest correlation is reported in parentheses. Effect size does not vary due to different pretest-posttest correlation, hence only one effect size value per study is reported.

**Table 5**

*Results of Moderator Analysis with Studies with Pretest-Posttest Design*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Moderator variables | *g* | *SE* | 95% CI | | *p* |
| Lower | Upper |
| Text levela | **−.51 (−.53-−.50)** | **.125 (.116-.132)** | **−.758 (−.762-−.756)** | **−.270 (−.307-−.238)** | **.000 (.000-.000)** |
| Low text levelb | **.45 (.45-−.46)** | **.160 (.151-.168)** | **.141 (.121-.161)** | **.767 (.753-.779)** | **.005 (.002-.007)** |
| High text levelb | −.04 (−.06-−.00) | .130 (.120-.138) | −.292 (−.335-−.240) | .216 (.208-.231) | .775 (.647-.968) |
| Both text levelsb | **−.67 (−.70-−.64)** | **.208 (.188-.226)** | **−1.076 (−1.084-−1.064)** | **−.260 (−.326-−.200)** | **.002 (.000-.004)** |
| Feedback representation |  |  |  |  |  |
| Graphicalb | **.34 (.31-.36)** | **.094 (.080-.107)** | **.156 (.153-.158)** | **.526 (.450-.572)** | **.000 (.000-.001)** |
| Numericb | −.19 (−.20-−.17) | .113 (.095-.129) | −.411 (−.453-−.360) | .032 (.013-.051) | .092 (.068-.119) |
| Highlightingb | **.52 (.51-.54)** | **.171 (.151-.190)** | **.188 (.162-.212)** | **.860 (.805-.908)** | **.002 (.001-.0048)** |
| Text-basedb | **−.77 (−.77-−.76)** | **.153 (.137-.168)** | **−1.069 (−1.093-**−**1.039)** | **−.468 (−.504-−.434)** | **.000 (.000-.000)** |
| Number of representationsc | .12 (.11-.13) | .204 (.202-.243) | −.316 (−.344-−.287) | .564 (.506-.610) | .580 (.577-.590) |
| Feedback specificityd | **.68 (.66-.71)** | **.206 (.187-.223)** | **.281 (.226-−.342)** | **1.089 (1.075-1.101)** | **.001 (.000-.003)** |
| Prior knowledgee | −.01 (−.01-−.01) | .013 (.011-.014) | −.033 (−.037-−.029) | .156 (.015-.017) | .477 (.456-.526) |
| Settingf | **−.82 (-.82-−.80)** | **.198 (.179-.215)** | **−1.204 (−1.244—1.155)** | **−.427 (−.452-−.401)** | **.000 (.000-.000)** |
| Educational levelg | **.73 (.71-.74)** | **.206 (.187-.223)** | **.325 (.277-.378)** | **1.133 (1.110-1.152)** | **.000 (.000-.000)** |

*Note.* Number of studies and effects = 24, total *N* = 2,380 students, *g* is the effect size of Becker’s (1988) standardized mean change from students’ first draft (pretest) to their revised draft (posttest), *SE* is the standard error of the effect size, 95% CI is a 95% confidence interval with its lower and upper limit. The table shows the mean effect size, its standard error, confidence interval, and the corresponding *p*-value across the 26 meta-analyses based on the mean correlation according to the sensitivity analysis procedure. The range of the respective values of the 26 meta-analyses with the smallest and largest correlation is reported in parentheses.

a 1 = feedback that addressed lower level of text quality only, 2 = feedback that addressed higher level of text quality only, 3 = feedback that addressed both lower and higher level of text quality. b 0 = no, 1 = yes. c 0 = mono (one single representation format), 1 = multiple representation formats combined. d 1 = generic, 2 = specific. e Students’ writing quality of their first draft (pretest) on a scale from 0% to 100%; it should be noted that the effect shown in the table is not a small decrease in knowledge, but *g* describes the size and direction of the moderation effect. f 1 = laboratory, 2 = field study. g 0 = school (including middle and high school), 1 = university.

**Table 6**

*Results of Moderator Analysis with Studies with Pretest-Posttest-Control Group Design*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Moderator variables | *g* | *SE* | 95% CI | | *p* |
| Lower | Upper |
| Text levela | **−.52 (−.52-−.52)** | **.090 (.077-.100)** | **−.717 (−.696-−.672)** | **−.345 (−.371-−.323)** | **.000 (.000-.000)** |
| Low text levelb | **.55 (.54-.56)** | **.103 (.085-.119)** | **.351 (.330-.376)** | **.755 (.711-.796)** | **.000 (.000-.000)** |
| High text levelb | **−.48 (−.49-−.46)** | **.102 (.084-.117)** | **−.679 (−.693-−.660)** | **−.280 (−.330−-.234)** | **.000 (.000-.000)** |
| Both text levelsb | **−.53 (−.54-−.52)** | **.267 (.254-.278)** | **−1.057 (−1.088-−1.019)** | **−.012 (−.025-.001)** | **.045 (.040-.0497)** |
| Feedback representationc |  |  |  |  |  |
| Graphicalb | .35 (.35-.36) | .248 (.222-.270) | −.131 (−.179-−.079) | .840 (.790-.880) | .152 (.109-.192) |
| Numericb | −.17 (−.18-−.17) | .350 (.337-.360) | −.860 (−.872-−.843) | .510 (.479-.539) | .617 (.589-.644) |
| Highlightingb | .13 (.13-.13) | .268 (.239-.293) | −.393 (−.445-−.336) | .658 (.602-.702) | .619 (.579-.661) |
| Text-basedb | .25 (.24-.27) | .323 (.316-.327) | −.378 (−.379-−.375) | .887 (.861-.909) | .430 (.415-.446) |
| Number of representationsc | .26 (.24-.28) | .244 (.220-.264) | −.215 (−.235-−.194) | .741 (.668-.801) | .281 (.280-.284) |
| Feedback specificityd | **.53 (.52-.54)** | **.267 (.254-.278)** | **.012 (.001-.025)** | **1.057 (1.019-1.088)** | **.045 (.040-.0497)** |
| Prior knowledgee | **−.05 (−.05-−.04)** | **.021 (.019-.023)** | **−.091 (−.098-−.081)** | **−.007 (−.007-−.007)** | **.021 (.021-.023)** |
| Settingf | .12 (.10-.14) | .336 (.329-.342) | −.543 (−.548-−.536) | .775 (.740-.806) | .730 (.693-.770) |
| Educational levelg | −.11 (−.14-−.09) | .339 (.331-.344) | −.779 (−.812-−.741) | .549 (.538-.557) | .736 (.690-.781) |

*Note.* Number of studies and effects = 15, total *N* = 1,829 students, *g* is the effect size of Becker’s (1988) standardized mean change from students’ first draft (pretest) to their revised draft (posttest), *SE* is the standard error of the effect size, 95% CI is a 95% confidence interval with its lower and upper limit. The table shows the mean effect size, its standard error, confidence interval, and the corresponding *p*-value across the 26 meta-analyses based on the mean correlation according to the sensitivity analysis procedure. The range of the respective values of the 26 meta-analyses with the smallest and largest correlation is reported in parentheses. Significant results are highlighted in bold letters.

a 1 = feedback that addressed lower level of text quality only, 2 = feedback that addressed higher level of text quality only, 3 = feedback that addressed both lower and higher level of text quality. b 0 = no, 1 = yes. c 0 = mono (one single representation format), 1 = multiple representation formats combined. d 1 = generic, 2 = specific. e Students’ writing quality of their first draft (pretest) on a scale from 0% to 100%; it should be noted that the effect shown in the table is not a small decrease in knowledge, but *g* describes the size and direction of the moderation effect. f 1 = laboratory, 2 = field study. g 0 = school (including middle and high school), 1 = university.

1. Table 3 shows the ranges of the effect size, its standard error and *p*-value depending on the pretest-posttest correlation within the scope of the sensitivity analysis. The results showed that the variations of the effect size *g* are minimal (only from the third decimal place) and do not change the result, which is why we can assume a robust effect and a precise estimate of the correlation. [↑](#footnote-ref-2)
2. Regarding the effects of the studies with pretest-posttest-control group design, the variations of *g* were larger, depending on the assumed pretest-posttest correlation of our sensitivity simulation (see Table 4). This finding suggests that the effect is not very robust across all 26 meta-analyses, but depends on the correlation. The result should therefore be treated with caution. [↑](#footnote-ref-3)