Educational Psychology Review (2023) 35:11
https://doi.org/10.1007/s10648-023-09745-1

META-ANALYSIS

A Meta-analysis of the Worked Examples Effect on Mathematics Performance

Christina Areizaga Barbieri[1] · **Dana Miller-Cotto**[2] · **Sarah N. Clerjuste**[1] · **Kamal Chawla**[1]

Accepted: 13 January 2023

Abstract
The current meta-analysis quantifies the average effect of worked examples on mathematics performance from elementary grades to postsecondary settings and to assess what moderates this effect. Though thousands of worked examples studies have been conducted to date, a corresponding meta-analysis has yet to be published. Exclusionary coding was conducted on 8033 abstracts from published and grey literature to yield a sample of high quality experimental and quasi-experimental work. This search yielded 43 articles reporting on 55 studies and 181 effect sizes. Using robust variance estimation (RVE) to account for clustered effect sizes, the average effect size of worked examples on mathematics performance outcomes was medium with $g=0.48$ and $p=0.01$. Moderators assessed included example type (correct vs. incorrect examples alone or in combination with correct examples), pairing with self-explanation prompts, and timing of administration (i.e., practice vs. skill acquisition). The inclusion of self-explanation prompts significantly moderated the effect of examples yielding a negative effect in comparison to worked examples conditions that did not include self-explanation prompts. Worked examples studies that used correct examples alone yielded larger effect sizes than those that used incorrect examples alone or correct examples in combination with incorrect examples. The worked examples effect yields a medium effect on mathematics outcomes whether used for practice or initial skill acquisition. Correct examples are particularly beneficial for learning overall, and pairing examples with self-explanation prompts may not be a fruitful design modification. Theoretical and practical implications are discussed.

Keywords Worked examples effect · Mathematics performance · Self-explanation · Geometry · Cognitive load theory · Mathematics learning · Incorrect examples

✉ Christina Areizaga Barbieri
barbieri@udel.edu

[1] University of Delaware, 113 Willard Hall Education Building, Newark, DE 19716, USA

[2] Kent State University, Kent, OH 44242, USA

Published online: 30 January 2023

In 2019, only 41% of eighth graders in the USA scored at or above proficient in mathematics (National Assessment for Educational Progress, NAEP 2020). Proficiency rates for the lowest performers showed significant declines since the 2017 National Assessment for Educational Progress (NAEP, 2020). Failure to develop foundational mathematics skills in the elementary and middle grades prevents college success (NMAP, 2008) and the pursuit of STEM careers (Shapka et al., 2006). Limitations in mathematics success can limit one's career opportunities (Sadler & Tai, 2007) and overall quality of life. Struggles with mathematics are not limited to the middle grades or the USA but are wide-ranging and international. For example, according to the 2019 Trends in International Mathematics and Science Study (TIMSS) results, only seven of the 58 countries assessed scored at the *High* benchmark and one at the *Advanced* benchmark on the 4th-grade mathematics achievement measure. The remaining 50 countries, which includes the USA, scored at either the *Intermediate* or *Low* benchmarks. Thus, it is crucial to identify strategies that improve mathematics learning that could be easily implemented in classroom settings across the globe. Research in educational psychology has revealed many useful instructional strategies to support mathematics learning with varying degrees of success (Booth et al., 2017, for a review). One such strategy that has been heavily studied internationally within the domain of mathematics learning is the use of *worked examples* (i.e., worked-out problem solutions). These are implemented by replacing some or all problem-solving practice with worked examples to study, and examples are sometimes paired with similar practice problems. Despite their widespread use, a meta-analysis on the worked examples effect on mathematics learning has yet to be published. Considering this, we conducted a meta-analytic review to examine the overall effect of worked examples on mathematics performance to understand the varying magnitude of its overall effect across studies.

Background

Originating with a classic study by Sweller and Cooper (1985), the worked examples effect suggests that having learners study worked-out problem solutions is more effective for learning than problem-solving practice alone in terms of reducing problem-solving errors and increasing performance on mathematics assessments. This seminal work served as a springboard for thousands of new studies on the worked examples effect (Booth et al., 2015a; Renkl, 2014; for reviews) with examinations of the effects of a wide range of variations (e.g., Sweller, 2012). These variations include but are not limited to type (e.g., correct, incorrect, faded), usage (e.g., initial knowledge acquisition vs. practice after instruction), delivery format (e.g., paper tasks, computerized), and target content (e.g., mathematics, engineering).

Despite the large body of work conducted on worked examples suggesting their use in classrooms, a 2007 Institute of Education Sciences' (IES) Practice Guide concluded that there was only *moderate* evidence that interleaving worked examples and problem-solving was effective for learning (Pashler et al., 2007). Most of the work cited in this guide was conducted in the mathematics domain, though science and computer engineering studies were also represented. It is important to note that, although the IES practice guides are funded by the US Department of Education, the studies within the guides are conducted internationally. Thus, although the recommendations are geared towards US practitioners and policy-makers, they can also be treated as general recommendations for classrooms globally. Recommendations in these guides led to practical implications with some curriculum developers incorporating examples into their curricula or standards (see CCSS-M, 2010 for an example).

Meanwhile, researchers continue to work on improving the effectiveness of worked examples by focusing on design aspects as well as how examples are used. In the fourteen years that followed the 2007 IES Practice Guide, research on worked examples yields over 18,000 results in a Google Scholar search, relative to about 16,000 results from 1985 to 2007 (at the time of writing) from studies around the world. Most of this work is relevant to mathematics learning as mathematics is a well-structured domain reliant on learning successful problem-solving strategies and procedures. In 2012 and in 2015, IES published two more Practice Guides focused specifically on mathematics learning. The first focused on Grades 4 through 8 during which the primary focus is arithmetic with some pre-algebraic content and the second focused on teaching algebra in particular. Both guides point to worked examples as effective instructional tools for varying reasons and with varying levels of supporting evidence. For example, the 2012 guide explains that there is *moderate* evidence that supports "exposing students to multiple problem-solving strategies" and "helping students recognize and articulate mathematical concepts and notation" using worked examples. More recently, the 2015 guide suggests that there is *minimal* evidence to support "using solved problems to engage students in analyzing algebraic reasoning and strategies" through worked examples paired with practice problems, finding only three studies that supported this claim (i.e., Booth et al., 2013; Carroll, 1994). These guides are useful for practitioners as they are written for a lay audience and with direct recommendations for practice. Yet, they are limited in that they do not provide a statistical quantification of such effects and use a more descriptive approach to summarize the quality of findings. Further, they focus primarily on supporting rather than conflicting evidence. Some reviews have also been published attempting to summarize the general effects of worked examples along with their variations (e.g., Booth et al., 2015a; Renkl, 2014). Still, despite this large body of work that has great variability in findings and practical implications for mathematics learning, a meta-analysis on the worked examples effect on mathematics learning has yet to be published. In light of this, we conducted a meta-analytic review to more conclusively quantify the overall effect of worked examples on mathematics performance across a wider age span of learners and to examine potential explanatory variables of the variability in such effects.

Theoretical Underpinnings

The worked examples effect is considered a prototypic example of *cognitive load theory* (CLT; Sweller, 1988). CLT explains that humans' cognitive capacity is limited in what it can process and traditional problem-solving practice (i.e., rote practice of a learned procedure) overloads that capacity for a novice learner, leaving insufficient room for the proper acquisition and development of new knowledge structures or schemata. When practicing new material, learners may study worked examples—problems shown with fully worked out solutions—prior to attempting to solve a similar problem. While students study these examples, they are not required to generate random processes to solve the problem shown or attend to irrelevant features. Instead, they can devote their cognitive effort to studying only the relevant aspects for solving the problem and the correct process to acquire and develop schemata for the problem-type studied (Sweller, 2006). The belief is that, after studying multiple worked examples, the student can extract the features of or principle for solving a problem that are common to other problems of that type, to assist in avoiding cognitive overload, leading to automation of problem-solving in that domain.

Earlier interpretations of CLT suggested that when individuals are solving a problem, much of their cognitive effort is devoted to a mean-ends analysis, or the current problem state (the problem) and the goal state (the solution). This process, particularly for novice learners, imposes greater cognitive load, which prevents the learner from solving the problem or gaining any additional knowledge. CLT suggests, in an effort to remedy this problem, that instructional designs should limit splitting cognitive resources during the learning phase and the problem-solving phase simultaneously (Renkl & Atkinson, 2010). The simultaneous focus would place extreme demands on cognitive processes and inhibit learning. To that end, worked examples allow the learner to focus on learning without the need to problem solve. Renkl (2014) presents a thorough overview of the theory of example-based learning.

The use of worked examples in improving learning in mathematics, in particular, has been supported to varying degrees in laboratory studies and classroom studies (see Booth et al., 2015a for a review specific to STEM learning). Prior work has shown that implementing worked examples increased learning in classroom work (Carroll, 1994) and, in some cases, through homework assignments (Miller-Cotto & Auxter, 2019; Ward & Sweller, 1990) in well-structured areas such as mathematics (Rittle-Johnson & Star, 2007; Zhu & Simon, 1987). Benefits of studying worked examples include improving students' ability to solve problems that are very similar to those they studied (Trafton & Reiser, 1993) as well as more difficult problems than those studied (e.g., Catrambone, 1996, 1998). Some studies have found that studying worked examples reduces learning time compared to more problem-solving practice (e.g., Schwonke et al., 2009) or increased learning in the same amount of time (e.g., Atkinson et al., 2003). Benefits of worked examples have been found in short-term classroom studies (e.g., Carroll, 1994) as well as longer-term studies both in traditional classrooms (e.g., Booth et al., 2015b) and computerized learning environments (e.g., Schwonke et al., 2007). Worked examples have also been used to promote flexibility in problem-solving by having students compare a pair of worked examples demonstrating the same problem solved in two different ways (Rittle-Johnson & Star, 2007; Star & Rittle-Johnson, 2009). Due to this great variation in the literature, we focus our meta-analytic review broadly on varying study environments (e.g., laboratory and classroom studies), with varying mathematics performance measures (e.g., percent accuracy on a test, number of errors made during problem-solving), and in varying modalities (e.g., paper-and-pencil tasks, computerized tutors).

Variations in Design and Use

Although some concrete recommendations have been made regarding the design of examples (e.g., Atkinson et al., 2000), their design and use vary extensively across studies along with the study design itself. Thus, summarizing the overall effectiveness of examples, as well as what explains variability in effects, is challenging. Booth et al. (2015a) review the literature on the effect of the worked examples on STEM learning more broadly but also highlight studies on the variants of worked examples that could be employed to yield different effects. For example, they summarize the usage of self-explanation prompts in supporting learning from worked examples as well as the use of both correct and incorrect examples to varying effects. Another common variation *across* studies is whether worked examples are used to help students practice skills learned previously in instruction or used to introduce learners to new information or content. Though there are many other variations that have been tested across laboratory and classroom studies as well as in the implementation of worked examples, we focus our attention on these three variations as they are most prevalent in the included studies.

Worked Example Correctness

Worked examples are frequently presented as having been completed by a fictional student and paired with a similar practice problem (e.g., McGinn et al., 2015). Most commonly, worked examples show a correctly worked-out solution. However, incorrectly worked-out solutions are also used and thought to draw students' attention to the underlying principle that makes a solution incorrect and refine problem-solving. Incorrect examples, either alone or in combination with correct examples, have been found to reduce misconceptions (Durkin & Rittle-Johnson, 2012) and promote learning concepts and procedures (Adams et al., 2014; Booth et al., 2015b). Incorrect worked examples are often used in combination with correct examples. For instance, Große and Renkl (2007) studied the effects of a combination of correct and incorrect examples compared to correct examples only. They found that students with high prior knowledge benefitted more from a combination of correct and incorrect examples, whereas those with low prior knowledge benefitted from the correct examples only materials. Durkin and Rittle-Johnson (2012) found that a comparison of correct and incorrect examples reduced misconceptions, increased discussion of correct concepts, and improved students' procedural skill and conceptual understanding of decimal magnitude compared to comparison of correct examples but only at *delayed* posttest. Booth et al. (2015b) found that a combination of correct and incorrect examples improved students' conceptual algebra knowledge more than traditional problem-solving practice at posttest. Adams et al. (2014) compared the effects of incorrect examples focused on place value to problem-solving in a computerized tutor and found that incorrect examples were more effective than problem-solving alone. Barbieri and Booth (2020) found that incorrect examples were more effective at improving equation-solving abilities than a problem-solving control, while two variations of correct examples did not outperform the control on this measure. However, all three example conditions outperformed the problem-solving control on a conceptually-based algebra measure with equal magnitude regardless of example-type. Findings from the above studies suggest that the combination of incorrect and correct examples produces improved performance on outcomes but there is clearly variability in results. The comparison of effects between correct worked examples and incorrect examples (either alone or in combination with correct examples) is one of the most studied example features in the literature. Yet results are still inconclusive on which is more effective overall. Thus, one important moderator of the worked example effect that we examine is example correctness, comparing effect sizes from correct example conditions to those that used incorrect examples either alone or in combination with correct examples.

Worked Examples Paired with Self-Explanation Prompts

Worked examples are often paired with self-explanation prompts. These are typically written prompts that ask students to explain a strategy or concept displayed in a worked example. Some work suggests that self-explanations increase the effectiveness of worked examples (Berthold et al., 2009; Catrambone & Yuasa, 2006; Renkl et al., 1998). Self-explanation prompts are suggested to work because they force learners to make their knowledge explicit by generating and integrating new knowledge with prior knowledge (Chi, 2013). A recent meta-analysis found that prompting learners to self-explain leads to small to moderate effects depending upon the mathematics outcome type (e.g., procedural knowledge, conceptual knowledge, transfer) immediately after intervention (Rittle-Johnson et al., 2017). However, this meta-analysis examined effect sizes for self-explanation prompts in general within a range of instructional contexts, not solely those paired with worked examples. Huang and Reiser (2012) found that studying worked examples paired with self-explanation prompts was more beneficial for learning than studying worked examples alone or with instructional explanations provided. Yet not all work within the worked examples literature supports their inclusion. For example, Barbieri et al. (2019) found that worked examples paired with self-explanation prompts were more effective than worked examples alone and as effective as those paired with visual signaling cues overall. However, when these two features were combined (self-explanation prompts plus visual signaling cues), there was no benefit over worked examples alone. The authors explain that this combination may provoke a redundancy effect, in which the learner is being presented with the same or similar information in two different forms leading to lowered effectiveness or even deleterious effects compared to effects of either feature alone. Further, students who displayed high prevalence of misconceptions at the start of the study benefitted more from worked examples supplemented with visual signaling cues than worked examples supplemented with explanation prompts or worked examples alone. Thus, though explanation prompts are generally viewed as a helpful and now almost standard addition to examples, there is still variability in their effectiveness. As such, we examine the inclusion of self-explanation prompts with worked examples as a moderator.

Timing of Implementation and Learner Prior Knowledge

An important consideration examined in prior research on worked examples is the timing of implementation during phases of learning. Two phase models related to learning from worked examples vary in how they describe the nature and progression of specific phases of skill acquisition along with how examples should be used (Anderson et al., 1997; VanLehn, 1996). However, both support the reliance on examples in the earlier stages of skill acquisition with the goal that they then be faded out and replaced with practice problems. In both phase models, the goal is for learners to rely upon examples to initially acquire and/or refine information on procedures and then to use these examples as a guide during practice until they are no longer needed because they are stored in long-term memory for future reference.

Thus, some research on the worked examples effect uses worked examples to *introduce* new content and assist learners in developing new schemata within a domain. However, other work uses worked examples *after* initial instruction as a form of practice with a goal of further refining and strengthening existing schemata. More specifically, consideration of the interaction between learners' prior knowledge and the effectiveness of examples is one of the most heavily studied topics within example-based research. The *expertise-reversal* effect highlights the importance of these considerations. As learners increase in expertise, instructional procedures or strategies designed to support novices become less effective and eventually even counterproductive which leads to negative or adverse effects on learning (Kalyuga et al., 2003, 2012). This has been suggested of the worked examples effect in particular in that replacing some practice problems with worked examples is more effective than problem-solving alone in the early stages of skill acquisition. Yet, as knowledge and expertise increase, practicing problem-solving can be more effective than studying examples (Sweller et al., 2019).

Though some work does support this general expertise-reversal effect within the worked examples literature, the relationship is not always apparent. Findings vary in both the presence of any effect as well as its direction and which learners benefit. For example, though Kalyuga et al. (2003) did demonstrate this effect, Carroll (1994) demonstrated the worked examples effect with both low and high achievers. Booth et al. (2015b) found that a combination of correct and incorrect examples was more effective for algebra learning than a problem-solving control for students with low prior algebra knowledge. Heemsoth and Heinze (2014) found that students with high prior knowledge who studied incorrect examples outperformed those who studied correct examples in fraction multiplication and division. Barbieri and Booth (2016) compared incorrect examples to correct examples and a problem-solving control and found that incorrect examples were more beneficial than correct examples or the problem-solving control but *only* for students with low prior algebra knowledge.

Yet Adams et al. (2014) found incorrect examples equally beneficial for those with low and high prior knowledge compared to a problem-solving control. One method for considering learner expertise is determining at which point examples are being introduced to learners and whether variations in the timing of their introduction yield variations in effect. That is, we can consider whether learners are being asked to study examples at the introduction of a new topic or whether examples are being used as a form of practice after instruction. Another way to quantify a potential expertise-reversal effect is to test aptitude-by-treatment interactions. Indeed, several studies take this approach (e.g., Barbieri & Booth, 2016; Booth et al., 2015b). In the current study, we consider the timing of implementation of examples (i.e., the general phase of skill acquisition) as one potential moderator of the effect. That is, we determine whether examples were used to introduce a new topic or for practice after instruction and assess these factors as statistical moderators.

We also consider the role of prior knowledge using a more descriptive approach by providing a summary of findings from the included studies that specifically address aptitude-by-treatment interactions. Due to the lack of standardized measures of prior knowledge across the included studies as well as a lack of any measure of prior knowledge (e.g., pretest only administered in 45% of the included effect sizes) across many of the included studies, we focus on this component of our analysis with a more qualitative description. Only twelve of the included manuscripts test for such interactions, and the conclusions made by such studies are considered in the results section.

Overview of the Current Study

Beyond the discrepant findings related to the potential moderators reviewed, one challenge in summarizing the effects of worked examples is that most research conducted on worked examples either lacks a control group entirely *or* the comparison group is another type of worked examples condition that tests a variation in the design rather than an overall effect of example (e.g., Barbieri et al., 2019). Though these types of studies are important for learning about the effects of variations in design as well as the underlying learning processes occurring during example-based learning, they do not allow easy conclusions to be made about general worked examples effects. That is, the lack of random assignment and a control group that does not include worked examples as part of the manipulation make it difficult to easily determine causality or make generalized claims about the body of work conducted to date. The goal of the current meta-analysis is to understand and quantify the overall effect of worked examples on mathematics performance in grade school and post-secondary settings and to assess what accounts for differences in this overall effect compared to a non-worked example-based control (e.g., a business-as-usual control or a problem-solving control). We include worked examples of instructional manipulations that are implemented with whole classrooms or individually as well as using paper-and-pencil tasks, textbooks, or computerized tutors. A major aim of this meta-analysis is to learn from only rigorous studies on the worked examples effect, focusing on experimental and quasi-experimental designs that have an appropriate control group (i.e., one that did not use examples) so that we may estimate the average effect and also gain insights into key aspects of these instructional materials that support mathematics performance. We assess moderators of the example effect across experimental and quasi-experimental studies on mathematics performance measures using meta-regression. We examine three potential moderators for which variations are most prevalent in the included studies: correctness of example, inclusion of self-explanation prompts with examples, and timing of implementation (i.e., whether examples were used during the introduction of a new topic or for practice after initial instruction). It is important to note that the experimental manipulation in these studies is specific to example or control condition, but not necessarily to the moderators examined (i.e., example correctness, inclusion of self-explanation prompts, or timing of administration). That is, some included studies experimentally manipulated the moderators assessed such as Barbieri and Booth (2016) who manipulated correctness by including an incorrect only, correct only, and problem-solving control. Yet this was not the case for all studies coded on that particular factor. For example, we see that McLaren et al., 2015 had only an incorrect condition vs. a problem-solving control. The same limitation holds for the other two moderators assessed. Thus, results from moderator analyses are descriptive rather than causal. To reduce publication bias, we review both published and grey literature.

Methods

Inclusion and Exclusion Criteria

We include studies conducted globally with participants ranging from elementary age to adulthood. This global search included published journal articles and grey literature including dissertations. However, to ensure the project team could code all necessary components for analyses, only studies written in English were included.

Our research questions provided the basis for inclusion and exclusion criteria, as well as.

the foundation for the coding and analysis of studies. We followed procedures in accordance with PICOS (PICOS: Participants, Intervention, Comparison, Outcome, Setting; Higgins & Green, 2013), endorsed by the Cochrane Collaboration, to demonstrate our target populations, types of interventions, comparison groups, outcomes, and settings of interest. Our target population was any student of mathematics learning from elementary age (kindergarten) through adulthood. The type of intervention we focused on was any worked examples material focused on mathematics skills. The comparison groups we focus on are worked examples for the treatment and any control group that does not include worked examples. This included a business-as-usual control or problem-solving control. Outcome measures targeted were any measures of mathematics performance including accuracy or errors made during problem-solving. The target settings were either laboratory or classroom studies. To align with our proposed PICOS framework, we used the exclusionary criteria described below:

1. The study is not in English: Studies that were not originally written in English were excluded during the first round of screening to ensure that the team could properly code the studies.
2. The study is not a worked examples study: Studies that did not include a worked examples manipulation.
3. Target content other than mathematics: Studies that were based on topics other than mathematics (e.g., chemistry, biology). Studies focused on statistics learning were excluded if their primary focus was on learning to conduct statistical analyses within software (e.g., running analyses in SPSS) or coding (e.g., writing R-script) rather than understanding and/or being able to calculate the underlying computations of an analysis (e.g., calculating measures of central tendency).
4. No mathematics performance measure: Studies that did not include a continuous mathematics performance measure (e.g., outcomes such as reaction time or surveys).
5. Qualitative study: Studies that did not provide any quantitative data (i.e., studies that relied on qualitative rather than statistical analyses).
6. No control group: Studies that did not have a control group which did not include worked examples. That is, some studies examined variations in worked examples, and so each condition included examples. To isolate the effects of worked examples, a non-worked examples control group is needed.
7. Random assignment not used: To ensure high study quality and increase our confidence in causal inferences made from our findings, we excluded all studies that did *not* use random assignment to condition either at the student or classroom level. This ensured that we only yielded experimental or quasi-experimental studies. If a study did not use random assignment but explained their assignment mechanism, we retained that study as quasi-experimental. This only occurred in two of the studies we included to be further described in the results section.
8. Worked examples used collaboratively: Studies that did not test the individual effects of worked examples in math, but instead tested the effects of worked examples when worked on collaboratively. Our unit of analysis is at the level of the student rather than groups of students.

Next, the two primary coders reviewed forward and backward citations for the included studies yielded from the primary searches using Google Scholar. The process for screening forward citations included searching each included article in Google Scholar and then clicking the "cited by" link to view all sources in Google Scholar that cited that particular article. Then, these abstracts were screened using the same aforementioned exclusionary criteria. That is, coders screened the abstracts of all manuscripts that cited any of the aforementioned studies coded for inclusion. Forward citation yielded one article appropriate for inclusion. The process for screening backwards citations included screening the abstracts of all of the cited articles listed in the reference sections of the included manuscripts. Backward citation review yielded four articles appropriate for inclusion. Numbers corresponding to each step in this exclusionary process are displayed in the flowchart in Fig. 1.

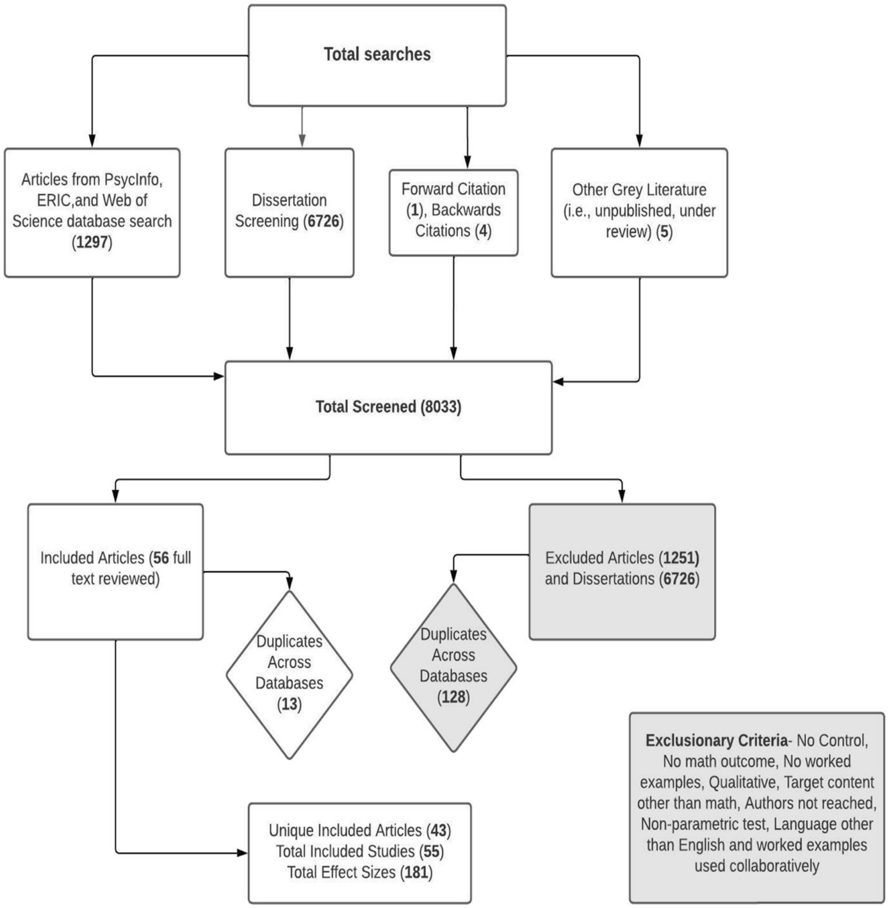


Fig. 1 Literature search and the process of inclusion and exclusion

Literature Search Procedure

The sample for the current meta-analysis includes worked examples studies that target mathematics learning. We performed an initial systematic search of ERIC, PsycINFO, ProQuest Dissertations, and Theses Global to identify all potentially relevant studies. We used the following search terms: worked examples AND math/mathematics; erroneous examples AND math/mathematics; incorrect examples AND math/mathematics; and faded worked examples AND math/mathematics. The ERIC and PsycINFO searches yielded an initial result of 763 items with 128 duplicates, while the ProQuest Dissertations and Theses Global yielded an initial result of 6726. Additional search processes for relevant literature included hand-searched reference sections (backward citations) of included articles, forward citations of included articles, and invitations for grey literature (i.e., unpublished) on an academic listserv (i.e., COGDEVSOC) and through Twitter. A second round of searches was conducted to add an additional database—Web of Science—and to add search terms. The additional pairs of search terms were chosen to ensure a broader definition of mathematics and included: worked examples AND statistics; erroneous examples AND statistics; incorrect examples AND statistics; and faded worked examples AND statistics. In this second round of searches, these terms were also searched in the databases used in the initial search (ERIC, PsycInfo, ProQuest Dissertations, and Theses Global) in addition to Web of Science. This second round yielded 534 abstracts to be screened. Figure 1 below provides a detailed flowchart of the included and excluded studies in the meta-analysis.

Exclusionary Coding

Exclusionary coding was completed by a team of coders using a combination of excel and Abstrackr (Wallace et al., 2012). Titles and abstracts were screened first. If the coder could not easily make an exclusionary decision based on the abstract, then the manuscript or dissertation itself was reviewed. After initial coding exclusion by one coder, another team member then re-coded for exclusion to ensure agreement. That is, all articles were screened for exclusion twice by two separate coders. Any discrepancies were resolved via discussion until the team reached 100% agreement on exclusion coding.

Data Extraction

Once the team agreed upon the articles for inclusion, coders began the data extraction phase. In this phase, each of the coders reviewed and extracted necessary information and data from each of the included articles. Information/data captured from each article included the means, standard deviations, and sample sizes that corresponded to each posttest given. We also coded for outcome type (e.g., accuracy on mathematics problem-solving measures, number of errors made on a measure), the students' grade (or age), how and by whom the study was administered (e.g., researcher, teacher, computerized tutor), the content targeted in the study (e.g., algebra, geometry), and the type of worked examples presented in the study (e.g., correct examples, incorrect examples). Additionally, we coded for whether worked examples were presented to students alone or paired with a practice problem. Finally, we coded for whether or not worked examples were paired with self-explanation prompts and whether worked examples were implemented during initial skill acquisition or for practice. All articles were coded by two independent coders. Initial percent agreement ranged from 91.40 to 100.00% depending on the particular source of data being coded (e.g., posttest means and standard deviations, sample sizes). Any discrepancies were resolved via discussion until the team reached 100% agreement on data extraction.

If any information needed to compute effect sizes was missing, authors were contacted via email to provide the relevant information. First authors who did not respond were contacted again a few weeks later with a second request. After the second request with no reply, a co-author was contacted. If there was still no reply at this point, then the article was excluded from the analysis. Two articles that were initially coded as include had to be coded as exclude due to missing descriptive statistics and unresponsive authors. This process yielded a final 43 total articles available for inclusion. Table 1 below provides key extracted information from the included articles.

Moderators

Some of the descriptive codes were then used to create dummy codes for moderator analyses described next. As aligned with our research questions, we assessed example type (correct vs. incorrect examples either alone or in combination with correct examples), self-explanation prompts (worked examples paired with explanation prompts or not), and phase of use (worked examples used during initial skill acquisition or practice after instruction). Moderators were chosen because they have been suggested in prior research to alter the effects of worked examples (e.g., incorrect examples; Barbieri & Booth, 2020) on learning.

One moderator of interest was prior knowledge. However, not all of the included studies used a pretest measure, and those that did include a measure did not always use a pretest identical to the corresponding posttest. Measures of prior knowledge that were included also necessarily varied across studies. Additionally, at the time of writing, there is no agreed upon method for statistically summarizing the effects of interaction terms across studies. Attempting to do so would not be feasible considering the great range in the number and types of conditions being assessed within each study and within each set of analyses as well as the variability in outcome measures. Thus, we take a descriptive approach to synthesizing findings from the literature that correspond to aptitude-by-treatment interactions. Specifically, two coders descriptively coded or summarized findings from the twelve manuscripts that tested aptitude-by-treatment interactions. Coders then compared these notes for consistency and summarized them for presentation in the results section.

Effect Size Calculations and Sampling Error Variance

Given our interest in the overall effect of an intervention, worked examples, on mathematics performance with a small sample of studies, we used the standardized mean difference, Hedge's *g*, as our effect size. Using the mean, standard deviation, and sample size of both the treatment and comparison/control group, we calculated Hedges' *g* on 181 effect sizes. The formula for Hedge's *g* is a proportion of the difference of the means over the pooled and weighted standard deviation (Hedges, 1981). The formula is as follows is as follows:

$$\text{Hedges'}g = \frac{M_1 - M_2}{SD^*_{pooled}}$$

Many of the articles and studies included in the current meta-analysis presented more than one effect size. Thus, we used Robust Variance Estimation (RVE) to

Table 1 Descriptions of included published studies

Study authors (year)	Country	Study design	N	Grade (or age)	Administered by/context	Content targeted	Worked example type(s)	Examples paired	Outcome type	Self-explanation prompts	Use
Adams et al. (2014)	USA	X	208	6th and 7th graders	T/C	Decimal magnitude	I	PP	A	E	P
Barbieri and Booth (2020)	USA	X	206	8th	T/C	Algebra	T, I	PP	A	E	P
Barbieri and Booth (2016)	USA	X	125	7th, 8th	T/C	Algebra	T, I	PP	A	E	P
Barbieri and Devlin, under review	USA	X	230	6th graders	T/C	Pre-algebra/ number line estimation	T,I,F	WE	A	N	P
Barbieri and Devlin, under review	USA	X	130	Grades 4 – 11	C/C	Fraction equivalence and magnitude	T, I,C	PP	A	E	P
Barbieri et al. (2021)	USA	X	75	8th grade	T/C	Algebra	T,I	PP	A, E	E	A
Begolli et al. (2021)	USA	X	220	7th grade	T/C	Probabilities and expected value	T,I,F	PP	A	E	P
Bentley and Yates (2017)	Australia	X	82	7th, 8th	T/C	Proportional reasoning	T	PP	A	N	P
Bokosmaty et al. (2015)	Australia	X	480	7th,8th,9th &10th	T, R/C	Geometry	T	PP	A	N	A
Booth et al. (2015c)	USA	X	380	8th	T/C	Algebra I	T,I	PP	A	E	A
Booth et al. (2013)	USA	X	198	High School	C/C	Algebra I	T, I	WE	A	E	P
Booth et al. (2015b)	USA	X	476	High School	T/C	Algebra I	T,I	PP	A	N	A

Table 1 (continued)

Study authors (year)	Country	Study design	N	Grade (or age)	Administered by/context	Content targeted	Worked example type(s)	Examples paired	Outcome type	Self-explanation prompts	Use
Carroll (1994)	USA	Q	24	High School	T/C	Algebra	T	PP	E	N	A
Chen (2016)	China	X	130	7th, 10th, 11th graders	T/C	Geometry, trigonometry	T	PP	A	N	A
Chen et al. (2015)	China	X	69	4th,7th	Not noted/P	Geometry	T	PP	A	N	P
Chen et al. (2019)	Indonesia	X	89	7th, 8th	Not noted/P	Algebra	T	WE	A	N	P
Chung & Tam (2005)	Hong Kong	X	30	8 to 12 years old	R/P	Arithmetic	T	PP	A	N	A
Cook (2017)	USA	X	260	Undergraduates	R/P	Statistics; correlations	T	PP	A	N	P
Cozean (2010)	USA	X	40	High School	R/P	Algebra	T	PP	A	N	A
Earnshaw (2014)	USA	X	143	6th – 8th	R/C	Proportional reasoning	T	WE	A	N	P
Fitzsimmons et al. (2021)	USA	X	66	1st and 2nd graders	C/P	Number line estimation	I, T	WE	A	N	A
Hartmann et al. (2021)	Germany	X	173	Secondary school students	C/C	Statistics	T	WE	A	N	A

Table 1 (continued)

Study authors (year)	Country	Study design	N	Grade (or age)	Administered by/context	Content targeted	Worked example type(s)	Examples paired	Outcome type	Self-explanation prompts	Use
Hilbert et al. (2008)	Germany	X	111	Student Teachers	T/C	Probability	C	WE	A	N	P
Hsiao et al. (2013)	Taiwan	X	107	Undergraduates	C/C	Business mathematics	T	PP	A	N	P
Likourezos et al. (2019)	Australia	X	103	Adults (age: 18 to 55)	R/C	Algebra	T	WE	A	N	P
McLaren et al. (2015)	USA	X	390	6th	T/C	Decimal magnitude and arithmetic	I	PP	A	E	P
Miller- Cotto (2017)	USA	X	114	Middle school	C/C	Geometry	F	PP	A	E, N	A
Miller-Cotto and Auxter (2019)	USA	X	135	Undergraduates	T/C	Algebra	F	PP	A	E, N	P
Mwangi and Sweller (1998)	Australia	X	18	3rd	R/C	Algebra	T,I	PP	A	E	A
Newman and DeCaro (2019)	USA	X	190	Undergraduates	R/L	Statistical variance	T	WE	A	N	P
Nye et al. (2018)	USA	X	103	College	C/L	Basic algebra	T	WE	A	E	P
Reed et al. (2013)	USA	X	128	High School	C/L	Algebra	T	PP	E	N	P
Reiss et al, (2008)	Germany	X	243	8th grade	T/C	Geometry	T	WE	A	E	P

Table 1 (continued)

Study authors (year)	Country	Study design	N	Grade (or age)	Administered by/context	Content targeted	Worked example type(s)	Examples paired	Outcome type	Self-explanation prompts	Use
Retnowati et al. (2017)	Indonesia	X	122	7th grade	T/C	Algebra	T	PP	A	N	A
Retnowati et al. (2010)	Indonesia	X	108	7th	T/C	Geometry theorems	T	PP	A	N	A
Salden et al. (2010)	Germany	X	57	High school	C/P	Geometry	F	PP	A	E	A
Schwonke et al. (2007)	Germany	X	80	8th and 9th graders	C/Not noted	Geometry	T	WE	A	E	A
Star et al. (2015)	USA	X	1367	8th,9th	T/P	Algebra I	C	PP	A	N	A
Ter Vrugte et al. (2017)	Netherlands	X	103	Prevocational school	R, T/P	Proportional reasoning	F	WE	A	N	P
van Harsel et al. (2019)	Netherlands	X	205	First-year college	R/P	Geometry	T	PP	A	N	A
van Loon-Hillen et al. (2012)	Netherlands	Q	45	4th	T/C	Arithmetic	T	WE	A	N	A
Zimmerman (2015)	USA	X	102	Undergraduates	C/P	Statistics	T	PP	A	N	P
Zormelo (2018)	South Africa	X	104	High school	R, T/C	Geometry	T	PP	A	N	A

Note. Study type: *X* experimental, *Q* quasi-experimental; administered by: *R* research, *T* classroom teacher, *C* computerized program; context: *C* classroom, *L* lab, *P* pull-out study. Example types: *T* traditional correct, *I* incorrect, *C* comparison, *F* faded. Study outcome: *A* accuracy, *E* number of errors. Self-explanation prompts: *E* explanations, *N* none. Use: *P* practice, *A* initial skill acquisition. Examples paired: *PP* paired with practice problem, *WE* worked example alone account for the nesting of effect sizes within manuscripts when calculating the average effect size. RVE uses variation in the effect sizes within the dataset to estimate standard errors rather than assuming variance from a model (Hedges et al., 2010). We report a random-effects model as we expect effect sizes for worked examples to vary across studies. A multivariate meta-analysis allowed us to account for dependency of multiple outcomes or subgroups within a particular article.

Heterogeneity

To measure heterogeneity in effect sizes, we calculated the I^2 statistic. The I^2 statistic describes the percentage of variation among studies that is due to true heterogeneity rather than sampling error or chance (Higgins & Thompson, 2002; Higgins et al., 2003). Cochran's Q is the traditional measure of heterogeneity that is calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies, with the weights determined by the pooling method. When the number of studies is small, as in most (and our) meta-analyses, Q has low power to detect heterogeneity (Gavaghan et al., 2000). Unlike Q, I^2 is not constrained by the number of studies included. The formula for I^2 is as follows:

$$I^2 = \frac{(Q - df)}{Q} \times 100$$

where Q is the chi-squared statistic and df is its degrees of freedom (Higgins & Thompson, 2002; Higgins et al., 2003).

Results

Study Descriptors

We included 43 articles that examined the worked example effect. The 43 articles reported a total of 55 studies which included a total of 181 effect sizes across all studies. Study descriptors with associated codes are in Table 1. Of these 55 studies, two used quasi-experimental methods. These two studies were presented in Carroll (1994) and van Loon-Hillen et al. (2012). Rather than using true random assignment, Carroll (1994) used matching by pretest scores to assign students to condition. Van Loon-Hillen et al. (2012) referred to their study as quasi-experimental, but did not explain why they chose that designation. They do state that they chose one classroom for the experimental condition and the other for the control, but no further rationale was provided. The remaining 53 studies used random assignment to condition. Also, 35 articles included studies which used correct worked examples. Thirteen articles included studies which used incorrect examples and three included comparison conditions. Six articles included studies that used faded example conditions. Most studies were conducted on middle school mathematics students and beyond, with only five studies conducted at the elementary school level. As such, most of the content targeted was either algebra or geometry. About half of the articles (23) were reports of studies conducted within the US and the remaining (20) conducted in countries outside of the USA (e.g., Germany, China, Australia). Teacher-administered studies were most common with twenty-two articles reporting on such studies. Eleven of the articles included researcher-administered studies, and ten reported on those administered within a computerized tutor (some multi-study articles reported more than one administration type). The majority of studies reported included performance outcomes measured by percent accuracy with just three articles reporting on problem-solving errors. Most notably, only four articles were published by or before 2007, which is the year that IES published its practice guide suggesting worked examples as a moderately effective tool for learning. The current analyses will serve to update our knowledge on the worked examples effect on mathematics learning, which has both practical and theoretical implications.

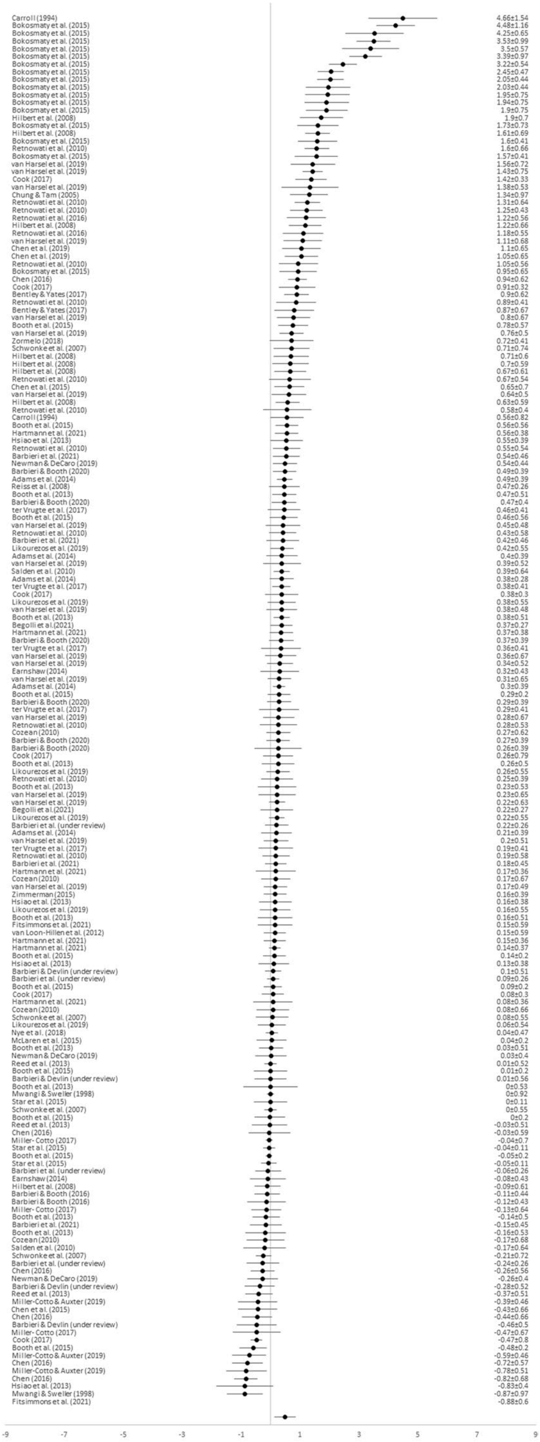
The Worked Example Effect on Mathematics Outcomes

To synthesize the 181 effect sizes across the 55 studies included, we examined the overall effect of the worked examples manipulations across studies. We yielded negative effect sizes for three articles (i.e., Barbieri et al., 2021; Carroll, 1994; Reed et al., 2013) that included number of problem-solving errors or misconceptions as outcomes which represented a benefit of worked examples over the control (i.e., less errors is an improvement). Thus, signs were reversed in these cases to represent the benefit or improvement. After this reversal, effect sizes ranged from −0.88 to 4.66. There were 38 effect sizes in the negative direction, four effect sizes with a value of zero and 139 in the positive direction. The average effect size was $g=0.48$ within our robust variance estimation (RVE) model with a 95% CI from 0.36 to 0.60, $p=0.01$, τ^2=0.5854. According to Cohen's standards, worked examples have a medium effect on mathematics outcomes. Significant heterogeneity was also detected, with I^2 = 93.72%. Effect sizes are displayed in the forest plot in Fig. 2.

Publication Bias

We examined potential publication bias through a series of publication bias tests. The funnel plot, illustrated in Fig. 3, indicates the relationship between study size and effect size. Here, study size refers to the standard error of the effect estimate and serves as a measure of study precision. Tests of funnel plot asymmetry examine whether the relationship between effect sizes and study size is greater than chance (i.e., that larger studies have markedly larger effect sizes). Inspection of the contour-enhanced funnel plots (see Fig. 3) demonstrated a relatively symmetric shape around the mean. To resolve any uncertainty, Egger's test was then used to examine the potential asymmetry of the funnel plot. Results of the Egger's test indicated a significant intercept, $B=-0.45$, $z=5.87$, 95% CI [−0.78, −0.12], $p<0.0001$, suggesting publication bias. However, subsequent trim-and-fill analyses showed that the effect of worked examples on mathematics performance would still be positive and significantly different from zero after adjusting for publication bias ($g=0.44$, SE = 0.06, 95% CI [0.32, 0.56], $p=0.0001$. Finally, the fail-safe N test using the Rosenberg

Fig. 2 A forest plot of effect sizes for all included studies. The circles represent the effect size for each study listed, the lines represent 95% confidence intervals, and the last circle represents the synthesised effect size



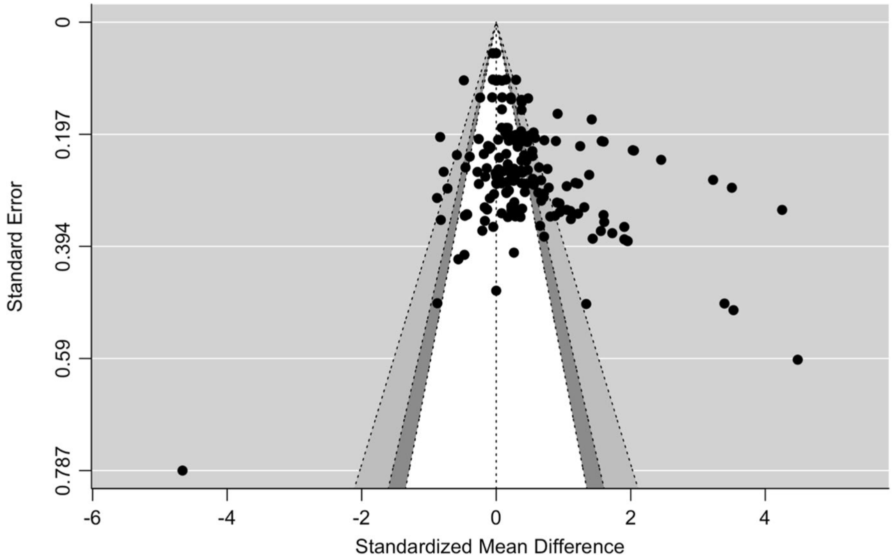


Fig. 3 Funnel plot of standard error by effect sizes

approach indicated that 13,915 studies would be needed to reverse the positive, significant effect size observed in the current study. Thus, no further publication bias tests were conducted.

Moderator Results

We examined three sets of moderators to determine what may explain some of the heterogeneity in effect sizes. Moderator model 1 examined example type by including a single dummy code for effect sizes from correct examples manipulations (1 = correct examples; 0 = incorrect examples) either alone or in combination with correct examples. Correct examples alone was the most common example-type with $n = 139$ effect sizes reflecting the effect of correct example conditions. The remaining 42 effect sizes corresponded to either incorrect examples alone ($n = 14$) or in combination with correct examples ($n = 28$). Moderator model 2 examined the inclusion of self-explanation prompts as a potential moderator. That is, a dummy code was created for effect sizes from example-based conditions that paired examples with self-explanation prompts. The referent in this model is effect sizes from example-based conditions that did not include self-explanation prompts. The majority of effect sizes were from worked examples conditions without self-explanation prompts ($n = 124$) with $n = 57$ effect sizes using self-explanation prompts along with examples. The third moderator model examined timing of administration as a potential moderator of effect sizes. One dummy code was created for effect sizes from studies that used worked examples for practice. The referent group was studies that used worked examples in the initial stages of skill acquisition as this was the larger of the two groups. That is, $n = 99$ effect sizes corresponded to worked examples conditions that were used in the initial stages of skill acquisition and $n = 82$ effect sizes corresponded to the use of worked examples for practice after instruction. Table 2 below presents findings from these models which show that although the example effect did not vary by timing of administration ($\beta = -0.15$; $p = 0.38$), it did indeed vary by correctness and whether examples were paired with explanation prompts. More specifically, correct worked examples ($\beta = 0.26$; $p = 0.03$) had a significant positive impact on the worked examples effect, and example-based conditions paired with self-explanation prompts had a significant negative impact on the worked examples effect, with ($\beta = -0.24$, $p = 0.04$).

The Role of Prior Knowledge in the Worked Examples Effect

Of the 43 manuscripts included in the current meta-analysis, twelve of them tested aptitude-by-treatment interactions. Each of these studies included an assessment of whether worked examples conditions varied in effectiveness for those with varying levels of prior knowledge. The nature of the worked examples as well as the comparison group varied by study. Some studies included several worked examples conditions, whereas others tested only one. Target content of examples also varied greatly as well as outcome measure. Thus, we approach the presentation of this summary descriptively rather than statistically.

Of the twelve manuscripts that assessed aptitude-by-treatment interactions, two did not yield any significant interaction (i.e., Carroll, 1994; McLaren et al., 2015). The remaining ten manuscripts present at least one significant interaction. Six of the twelve manuscripts focus on algebra learning with the remaining focusing on other areas of mathematics such as geometry and rational number knowledge. Of the twelve manuscripts, four manuscripts use correct examples only. Two of these four focus on faded examples and self-explanation prompts as key design features (i.e., Miller-Cotto, 2017; Miller-Cotto & Auxter, 2019). The remaining eight manuscripts include worked examples conditions that present incorrect worked examples either alone or in combination with correct examples. We summarize the findings below.

Four manuscripts found that the effectiveness of varying forms of worked examples did not vary based on prior knowledge or achievement (i.e., Carroll, 1994; McLaren et al., 2015; Miller-Cotto, 2017; Miller-Cotto & Auxter, 2019). These

Table 2 Moderator models of effects of worked examples on mathematics

	Example type model			Self-explanation model			Practice model		
Moderators	B	*SE*	*p*	β	*SE*	*p*	B	*SE*	*p*
Correct examples	0.26*	0.10	0.027	-	-	-	-	-	-
Paired with self-explanation prompts	-	-	-	-0.24*	0.11	0.042	-	-	-
Used for practice	-	-	-	-	-		−0.15	0.16	0.378

*$p < 0.05$ level

particular studies used a variety of example types including correct examples (i.e., Carroll, 1994), faded examples with and without self-explanation prompts (i.e., Miller-Cotto, 2017; Miller-Cotto & Auxter, 2019), and incorrect examples (i.e., McLaren et al., 2015) across a range of mathematics topics (e.g., geometry, algebra, rational number knowledge).

However, eight manuscripts did indeed find that the effectiveness of examples varied by prior knowledge. Yet the direction of these effects varied. Three studies found particular benefits for those with high prior knowledge in the domain (i.e., Bokosmaty et al., 2015; Fitzsimmons et al., 2021; Hartmann et al., 2021), but this was specific to example type in the study by Bokosmaty and colleagues. That is, Bokosmaty and colleagues found that expert geometry learners benefitted more from an example condition with step guidance, whereas novice learners benefitted more from an example condition with both step and theorem guidance. Fitzsimmons et al. (2021) found that participants with high prior knowledge (i.e., those more accurate in their number line estimations at pretest) benefitted more from correct examples paired with feedback than incorrect examples or a problem-solving control. These findings are in contrast with Hartmann et al. (2021) who found that students with high prior knowledge benefitted more from an incorrect worked examples condition than a problem-solving control condition. Both Fitzsimmons et al. (2021) and Hartmann et al. (2021) are in contrast with findings from the remaining five studies. A series of studies by Booth, Barbieri, and colleagues all found particular benefits for those with low prior knowledge (i.e., Barbieri & Booth, 2016, 2020; Barbieri et al., 2022; Booth et al., 2015b, c). What these five studies have in common is that the example conditions included incorrect examples either alone or in combination with correct examples.

These results suggest that although several studies support the notion that worked examples conditions that incorporate incorrect examples either alone or in combination with correct examples is particularly beneficial for those with low prior knowledge in the target content (e.g., Barbieri & Booth, 2016, 2020; Barbieri & Devlin, under review; Booth et al., 2015b, 2015c), this finding is in no way conclusive. Several studies that include incorrect examples in some form do not find them to be especially beneficial for those with low prior knowledge (i.e., Fitzsimmons et al., 2021; Hartmann et al., 2021; McLaren et al., 2015).

Discussion

Worked examples have a positive and moderate effect on students' mathematics learning across the school years. Learners who study worked examples improve almost half of a standard deviation ($g = 0.48$) in their mathematics performance compared to those who do not. Only four of the 43 articles reviewed were published by 2007 when the IES published a report claiming moderate evidence to support interleaving problem-solving practice with worked examples. Our results confirm and extend these findings as well as those reviewed in Booth et al. (2015a). A potential explanation for these findings has been posed in prior literature on worked examples: that studying worked-out problem solutions encourage learners to devote their cognitive efforts to studying the relevant features and processes of solving the problem eventually leading to the development of schemata for that problem-type which improves later performance on similar problems (Sweller, 2012). Problem-solving alone is not as effective in accomplishing this goal.

As evidenced by our heterogeneity index, we observed a significant amount of variation in the effect sizes across studies in the current review. To address these, we examined two moderators based on cognitive load theory (Sweller, 2012): self-explanation and use (practice or acquisition). We examined one additional moderator: correctness of example. We explained some of this heterogeneity using two significant moderators, self-explanation prompts ($\beta = -0.24$) and correct examples ($\beta = 0.26$), both with small-medium effect sizes but one in the negative direction and the other in the positive. That is, though correct examples seem to be *more* effective than incorrect examples alone or in combination with correct examples, example conditions (across type) that include self-explanation prompts seem to be *less* effective than example conditions without self-explanation prompts.

The findings related to self-explanation prompts are somewhat surprising as some prior work that estimates the effect of self-explanation prompts suggests them as a useful addition to example-based materials. For example, both Barbieri et al. (2019) and Huang and Reiser (2012) experimentally assessed the effect of adding self-explanation prompts to example-based materials and found benefits to such an addition in comparison to examples alone. One possible explanation for the negative overall moderation by self-explanation could be that the efficacy of self-explanation prompts may vary depending on the type of prompt used (Dunlosky et al., 2013). For example, Kuhn and Katz (2009) asked students to engage in generated self-explanations that asked students to make claims about the causes of earthquakes and compared performance of these students to those who were not prompted. Results suggested that children prompted to self-explain performed *worse* than those without prompting. They explain that self-explanations were often not high quality which may suggest that children were unable to integrate novel information with prior beliefs, suggesting a role of prior knowledge. Though this example is outside of the mathematics domain, these findings point to potential explanations for the negative impact that explanation prompts had on the overall example effect size. High-quality self-explanations that are associated with strong learning outcomes often include principles (i.e., citing an underlying rule for why an answer must be accurate; Renkl, 2002); poor self-explanations—which do not include much detail, do not attempt to anticipate upcoming steps in a problem, or fail to demonstrate a clear understanding of the goal of a problem—are associated with worse learning (Matthews & Rittle-Johnson, 2009). Thus, it could be likely that prompts vary in the quality of the self-explanations evoked which then leads to varying effectiveness of their use. Additionally, in some instances, when combining self-explanation prompts with other design features, if the information is duplicated, this may create a redundancy effect as well (e.g., Barbieri et al., 2019; Sweller, 2012). Unfortunately, at the time of writing, most articles we reviewed did not provide many details about the types of self-explanation prompts included, preventing us from further coding the types of self-explanation prompts. This may be a fruitful area for future research:

differentiating between the features of prompts that make them more or less effective for student learning.

It is important to note that findings from moderator analyses are *not causal* primarily because we estimate the effect of these moderators by collapsing across worked examples conditions that may or may not be isolating effects of such features. For example, in some studies, the worked examples condition with self-explanation prompts is the only example condition in that particular study, with the comparison being the control group (e.g., often problem-solving alone). Thus, we are generally comparing effect sizes of worked examples conditions with self-explanation prompts to worked examples conditions without prompts *across* studies, rather than *within* study, and this type of comparison does not involve a crucial feature for making causal claims: random assignment to condition. Further, a larger proportion of effect sizes corresponded to worked examples conditions without explanation prompts (69%) versus with prompts (31%) making the comparison somewhat unbalanced. The same cautionary note applies to findings related to the correctness of example and the practice vs. acquisition results. That is, though example type was often experimentally manipulated within studies (e.g., Adams et al., 2014; Barbieri & Booth, 2020), not all studies used included in the moderator analyses for correct examples do so. Further, the majority of effect sizes related to correct example conditions (77%) compared to incorrect examples alone or in combination with correct examples (23%).

Relatedly, we were limited in some example types we had planned to assess separately (e.g., incorrect examples alone vs. in combination with correct examples; faded worked examples) and thus needed to collapse down to two types (i.e., correct examples only vs. incorrect either alone or in combination with correct examples). If a larger mass of studies is conducted on each of these example types, separate estimates of these effects may be calculated and compared more confidently in future work. Further, timing of administration was also not experimentally manipulated in any of the included studies, with the assessed worked examples condition in each study *either* used for practice *or* skill acquisition but not with a comparison of these two modes of use assessed within a single study. It may be the case that the reason for using worked examples, whether it is for practice or skill acquisition, matters less than the design features used in combination with worked examples. Still, experimental work is needed to causally assess the potential differences that may result from varying use of examples.

The findings related to correctness of example are somewhat surprising as much work examining the difference between the effectiveness of correct and incorrect examples either does not find a difference (e.g., Barbieri & Booth, 2016) or finds a benefit of incorrect examples (e.g., Barbieri & Booth, 2020). However, as our synthesis on the aptitude-by-treatment interactions revealed, the main effect of correctness collapsed across levels of prior knowledge may vary substantially from their effectiveness once we take into account learner prior knowledge in the estimates. Findings from this synthesis indicated a wide range of scenarios. Strong conclusions cannot be made due to the small number of studies that tested aptitude-by-treatment interactions but also due to the variations in example conditions assessed. Two general conclusions may be suggested from this synthesis. First, based on findings from two studies, faded examples do not seem to vary by prior knowledge. Second, based on findings from five studies, incorrect examples either alone or in combination with correct examples *may* be particularly beneficial for students with low prior knowledge. However, two studies did not find this same pattern of results.

It is possible, however, that correct examples are indeed more beneficial overall (as suggested by the current results) but that incorrect examples alone or in combination with correct examples may be more effective for those with low prior knowledge. It is also possible that these overall effects could vary depending on our outcome of interest. As the main goal of using worked examples is typically to build learners' schemata for more successful problem-solving which could then be displayed most directly on a performance measure (e.g., the types of posttests often used in such studies), it is possible that correct examples may help learners reach this goal more effectively than incorrect examples. That is, if the outcome we are using is general performance on a problem-solving task, it is understandable that a series of correct worked examples may help support building and refinement of correct problem-solving strategies to a greater degree than incorrect examples alone or in combination with correct examples. However, it is quite possible that if we alter the outcome of interest to something like reduction of misconception strength or increases in learners' negative knowledge (e.g., knowing what doesn't work, what not to do, etc.; Gartmeier et al., 2008), we may find either no difference between examples of varying correctness or a greater effect of incorrect examples. However, the vast majority of manuscripts (i.e., 93%) included in the current estimation of the average effect of examples used performance measures with percent accuracy as their sole target outcome. Thus, conclusions on the effects of example type (e.g., varying by correctness) on other forms of important mathematical skills are not possible with the current data. Further work is needed to make stronger conclusions about the role of prior knowledge in the effectiveness of varying example types as well as the effect of examples on varying outcome types.

Strengths and Limitations

The current meta-analysis has several methodological strengths of note. First, to enhance the strength of our causal inference our exclusionary criteria ensured, we only assessed experimental or quasi-experimental studies. This search yielded a majority of experimental studies with only two of the 43 articles included being quasi-experimental. We are also confident that our search process yielded all relevant studies (written in English), as the number of abstracts screened from grey literature was substantially greater than those published, reviewing 6726 dissertation abstracts and approximately 1300 published works. We yielded a large number of effect sizes (i.e., 181) across 55 studies within 43 articles. Finally, we employ robust variance estimation to appropriately handle the nesting of effect sizes. This is particularly important since many of the studies in the worked example literature tend to be nested within research groups as well. This technique allows us to further isolate such effects that may be relying on the same research teams.

Despite these strengths, the current analysis has several weaknesses in addition to the limitations already noted about moderator analyses. For example, some study characteristics we had hoped to quantify were not possible due to limitations in reporting details. We had hoped to quantify the potential effect of prior knowledge on worked examples. Unfortunately, this was difficult for two reasons. First, the variability in pretests made it difficult to compare these across studies. Second, the total number of studies that used a pretest, that number being the minority of total studies, was challenging. Such an analysis is a problem with meta-analysis as a statistical technique more generally. Thus, we were only able to provide a descriptive synthesis of these findings. Further, though we had initially hoped to assess whether the dosage of worked examples moderated their effect, we were unable to do so as most articles lacked specific measurements of time to incorporate into our coding and analysis. There is great variability in the likely dosage of worked examples within our sample of studies, ranging from less than one class period (e.g., ~ 25 min) to examples administered throughout the academic year. Those that did report a general timing or dosage of examples reported a low dosage of examples (e.g., one class period or a handful of class periods). It is likely that a greater dosage of examples would yield a larger effect. More specific reporting would assist in the future analysis of the impact of dosage on the worked example effect. It is possible that with the Open Science movement gaining popularity that more materials will become available for such coding purposes.

Finally, a limitation of the current review is that approximately half of the studies included in this meta-analysis were conducted in the USA. Though we did not intentionally exclude international studies, one of our exclusionary criteria was that the study is written in English to allow us to extract data from the manuscripts. Thus, most of our work is from the USA and English-speaking countries. This is a limitation of the process inherent when dealing with language barriers and can potentially be improved upon in future work by multilingual research teams.

Implications for Theory and Practice

Cognitive load theory acknowledges that there are several aspects of problems that may increase their difficulty: presentation (e.g., side-by-side for comparison vs. one at a time), element interactivity (i.e., the number of elements in a given problem), and low learner expertise or prior knowledge. Here, we examined two relevant moderators to cognitive load theory: the inclusion of self-explanation prompts and whether worked examples were used for practice or skill acquisition. For self-explanation prompts, more research providing direction on designing effective self-explanation prompts and in which forms they can be effective based on cognitive load theory would make a significant contribution to the literature. Future work experimentally manipulating correctness type may help to clarify findings related to the correctness moderator. This work may support the usage of correct examples alone compared to incorrect examples alone or in combination with correct examples when the target outcome is problem-solving accuracy in particular. Cognitive load theory does not currently support strong conclusions as to why learning from correct examples might be more favorable for students than incorrect examples. Future theoretical work on CLT should consider this feature as well as whether recommendations of CLT in relation to correctness of example may vary by learner expertise and target outcome.

The current meta-analysis provides additional evidence that worked examples alone or paired with similar practice problems remain a beneficial tool for teaching mathematics across a wide range of mathematics domains, though it is important to note that most manuscripts (60%) are related to algebra or geometry. More widespread use of worked examples in mathematics classrooms as well as professional development for teachers in how to create and use their own worked examples based on their students' needs may be a fruitful practical next step. It is important to note that the current findings are specific to the domain of mathematics. Worked examples are also examined in other STEM-related academic domains such as engineering and computer science though these are fewer in number than studies focused on mathematics learning. Future work may examine the average effect of examples in other domains to make practical recommendations and theoretical distinctions in effects of example.

Conclusions

The current meta-analysis confirms that studying worked examples alone or paired with similar practice problems are moderately effective at improving learners' mathematics performance across the school years, compared to solving practice problems alone. Most manuscripts reviewed in our meta-analysis were conducted in schools (91%) rather than laboratories (7%), although some were pull-out studies (26%) and others were situated in the classroom context (65%). This was not a formal exclusionary criterion but just a natural result of the search. However, this feature increases the ecological validity of our findings making them particularly applicable to the classroom. These study features and findings suggest that worked examples in their various forms are a useful and effective tool to implement in mathematics classrooms, with some forms (i.e., correct examples) being potentially more effective than others (e.g., incorrect examples) and some supplemental features (i.e., self-explanation prompts) being not necessary or even detrimental to their effectiveness of student performance. In a school content area that is often approached with great difficulty by so many learners, it is imperative that we carefully consider research findings when making suggestions to practitioners on tools to use to support learning. The current meta-analysis is a necessary step in consolidating the immense amount of research on the topic with the goal of providing more conclusive suggestions to practitioners on use of examples to support their students' mathematics performance.

Acknowledgements The authors would like to thank Lauren DeLuca, Jenna Landy, Phuc Le, Georgia McKinney, Emily O'Hara, and Lucia O'Neill for assisting in the coding for this project.

Data Availability Data are available upon request from the corresponding author.

Declarations

Conflict of Interest The authors declare no competing interests.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Adams, D. M., McLaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., & Van Velsen, M. (2014). Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior, 36*, 401-411. https://doi.org/10.1016/j.chb.2014.03.053

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 932–945. https://doi.org/10.1037/0278-7393.23.4.932

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70*(2), 181–214. https://doi.org/10.3102/00346543070002181

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology, 95*(4), 774–783. https://doi.org/10.1037/0022-0663.95.4.774

*Barbieri, C., & Booth, J. L. (2016). Support for struggling students in algebra: Contributions of incorrect worked examples. *Learning and Individual Differences, 48*, 36–44. https://doi.org/10.1016/j.lindif.2016.04.001

Barbieri, C. A., Miller-Cotto, D., & Booth, J. L. (2019). Lessening the load of misconceptions: Design-based principles for algebra learning. *Journal of the Learning Sciences, 28*(3), 381–417. https://doi.org/10.1080/10508406.2019.1573428

*Barbieri, C. A., & Booth, J. L. (2020). Mistakes on display: Incorrect examples refine equation solving and algebraic feature knowledge. *Applied Cognitive Psychology, 34*(4), 862–878. https://doi.org/10.1002/acp.3663

*Barbieri, C.A., Booth, J.L. & Chawla, K. (2022). Let's be rational: Worked examples supplemented textbooks improve pre-algebra students' conceptual and fraction magnitude knowledge. *Educational Psychology,* 1–21. https://doi.org/10.1080/01443410.2022.2144142

*Barbieri, C., Booth, J.L., Begolli, K.N., & McCann, N.F. (2021). The effect of worked examples on student learning and error anticipation in algebra. *Instructional Science, 49*(4), 419–439. https://doi.org/10.1007/s11251-021-09545-6

*Barbieri, C.A., & Devlin, B. (under review). Targeting fraction misconceptions and reducing high confidence errors: Effects of worked example types and problem-solving with corrective feedback.

*Begolli, K. N., Dai, T., McGinn, K. M., & Booth, J. L. (2021). Could probability be out of proportion? Self-explanation and example-based practice help students with lower proportional reasoning skills learn probability. *Instructional Science*, *49*(4), 441–473. https://doi.org/10.1007/s11251-021-09550-9

*Bentley, B., & Yates, G. (2017). Facilitating proportional reasoning through worked examples: Two classroom-based experiments. *Cogent Education, 4*(1), 1-14. https://doi.org/10.1080/2331186X.2017.1297213

Berthold, K., Eysink, T. H., & Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science, 37*(4), 345–363. https://doi.org/10.1007/s11251-008-9051-z

*Bokosmaty, S., Sweller, J., & Kalyuga, S. (2015). Learning geometry problem solving by studying worked examples: Effects of learner guidance and expertise. *American Educational Research Journal, 52*(2), 307–333. https://doi.org/10.3102/0002831214549450

Booth, J. L., McGinn, K. M., Barbieri, C., Begolli, K. N., Chang, B., Miller-Cotto, D., Young, L. K., & Davenport, J. L. (2017). Evidence for cognitive science principles that impact learning in mathematics. In D. C. Geary, D. B. Berch, R. Ochsendorf, & K. M. Koepke (Eds.), *Acquisition of complex arithmetic skills and higher-order mathematics concepts* (pp. 297–325). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-805086-6.00013-8

*Booth, J. L., Lange, K. E., Koedinger, K. R., & Newton, K. J. (2013). Using example problems to improve student learning in algebra: Differentiating between correct and incorrect examples. *Learning and Instruction, 25*, 24–34. https://doi.org/10.1016/j.learninstruc.2012.11.002

Booth, J. L., McGinn, K. M., Young, L. K., & Barbieri, C. (2015a). Simple practice doesn't always make perfect: Evidence from the worked example effect. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 24–32. https://doi.org/10.1177/2372732215601691

*Booth, J. L., Cooper, L., Donovan, M. S., Huyghe, A., Koedinger, K. R., & Paré-Blagoev, E. J. (2015b). Design-based research within the constraints of practice: AlgebraByExample. *Journal of Education for Students Placed at Risk, 20*(1–2), 79–100. https://doi.org/10.1080/10824669.2014.986674

*Booth, J. L., Oyer, M. H., Paré-Blagoev, E. J., Elliot, A., Barbieri, C. Augustine, A. A., & Koedinger, K.R. (2015c). Learning algebra by example in real-world classrooms. *Journal of Research on Educational Effectiveness, 8*(4), 530–551. https://doi.org/10.1080/19345747.2015.1055636

*Carroll, W. M. (1994). Using worked examples as an instructional support in the algebra classroom. *Journal of Educational Psychology, 86*(3), 360–367. https://doi.org/10.1037/0022-0663.86.3.360

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(4), 1020–1031. https://doi.org/10.1037/0278-7393.22.4.1020

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology, 127*(4), 355–376. https://doi.org/10.1037/0096-3445.127.4.355

Catrambone, R., & Yuasa, M. (2006). Acquisition of procedures: The effects of example elaborations and active learning exercises. *Learning and Instruction, 16*(2), 139–153. https://doi.org/10.1016/j.learninstruc.2006.02.002

*Chen, O. (2016). *The worked example effect, the generation effect, and element interactivity* [Unpublished doctoral dissertation]. University of New South Wales.

*Chen, O., Kalyuga, S., & Sweller, J. (2015). The worked example effect, the generation effect, and element interactivity. *Journal of Educational Psychology, 107*(3), 689–704. https://doi.org/10.1037/edu0000018

*Chen, O., Retnowati, E., & Kalyuga, S. (2019). Effects of worked examples on step performance in solving complex problems. *Educational Psychology, 39*(2), 188–202. https://doi.org/10.1080/01443410.2018.1515891

Chi, M. T. (2013). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in Instructional Psychology* (pp. 161–238). Routledge.

*Chung, K. K. H., & Tam, Y. H. (2005). Effects of cognitive-based instruction on mathematical problem solving by learners with mild intellectual disabilities. *Journal of Intellectual & Developmental Disability, 30*(4), 207–216.

Common Core State Standards Initiative. (2010). *National governors association center for best practices and council of chief state school officers*. Retrieved December, 11, 2012.

*Cook, M. A. (2017). *A Comparison of the effectiveness of worked examples and productive failure in learning procedural and conceptual knowledge related to statistics*. [Doctoral Dissertation, The Pennsylvania State University].

*Cozean, K. R. (2010). *Enhancing performance in high school algebra: Using guided homework as an instruction strategy* [Doctoral Dissertation, Capella University]. ProQuest Dissertations Publishing.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58.

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction, 22*(3), 206–214. https://doi.org/10.1016/j.learninstruc.2011.11.001

*Earnshaw, Y. (2014). *Effects of levels of instructional assistance on learning and mental effort in an intelligent tutoring system: Proportional reasoning and middle school students* (Publication No. 3637974) [Doctoral Dissertation, Florida State University]. ProQuest Dissertations & Theses A&I.

*Fitzsimmons, C. J., Morehead, K., Thompson, C. A., Buerke, M., & Dunlosky, J. (2021). Can feedback, correct, and incorrect worked examples improve numerical magnitude estimation precision? *The Journal of Experimental Education*, *91*(1), 20–45. https://doi.org/10.1080/00220973.2021.1891009

Gartmeier, M., Bauer, J., Gruber, H., & Heid, H. (2008). Negative knowledge: Understanding professional learning and expertise. *Vocations and Learning, 1*(2), 87–103. https://doi.org/10.1007/s12186-008-9006-1

Gavaghan DJ, Moore AR, McQuay HJ. (2000). An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. *Pain, 85*(3), 415–424. https://doi.org/10.1016/S0304-3959(99)00302-4

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction, 17*(6), 612–634.

*Hartmann, C., van Gog, T. & Rummel, N. (2021). Preparatory effects of problem solving versus studying examples prior to instruction. *Instructional Science*, *49*(1), 1–21. https://doi.org/10.1007/s11251-020-09528-z

Hedges L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128.

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

Higgins, J. P., & Green, S. (2013). Cochrane handbook for systematic reviews of interventions version 5.1. 0. Chapter 8: Assessing risk of bias in included studies. *The Cochrane Collaboration*, 2011.

Heemsoth, T., & Heinze, A. (2014). The impact of incorrect examples on learning fractions: A field experiment with 6th grade students. *Instructional Science, 42*(4), 639–657. https://doi.org/10.1007/s11251-014-9313-x

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

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

*Hilbert, T. S., Renkl, A., Kessler, S., & Reiss, K. (2008). Learning to prove in geometry: Learning from heuristic examples and how it can be supported. *Learning and Instruction, 18*(1), 54-65. https://doi.org/10.1016/j.learninstruc.2006.10.008

*Hsiao, J., Hung, C., Lan, Y., & Jeng, Y. (2013). Integrating worked examples into problem posing in a web-based learning environment. *Turkish Online Journal of Educational Technology, 12*(2), 166–176

Huang, X., & Reiser, R. A. (2012). The effect of instructional explanations and self-explanation prompts in worked examples on student learning and transfer. *International Journal of Instructional Media, 39*(4), 331–344.

Institute of Education Sciences (IES). (2007). *WWC: Organizing instruction and study to improve student learning*. Retrieved December 26, 2021. https://ies.ed.gov/ncee/wwc/practiceguide/1

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., Rikers, R., & Paas, F. (2012). Educational implications of expertise reversal effects in learning and performance of complex cognitive and sensorimotor skills. *Educational Psychology Review, 24*(2), 313–337. https://doi.org/10.1007/s10648-012-9195-x

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology, 103*(3), 386–394.

*Likourezos, V., Kalyuga, S. & Sweller, J. (2019). The variability effect: When instructional variability is advantageous. *Educational Psychology Review, 31*(2), 479–497. https://doi.org/10.1007/s10648-019-09462-8

Matthews, P., & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology, 104*(1), 1–21.

McGinn, K. M., Lange, K. E., & Booth, J. L. (2015). A worked example for creating worked examples. *Mathematics Teaching in the Middle School, 21*(1), 26–33. https://doi.org/10.5951/mathteacmiddscho.21.1.0026

*McLaren, B.M., Adams, D.M. & Mayer, R.E. (2015). Delayed learning effects with erroneous examples: A study of learning decimals with a web-based tutor. *International Journal of Artificial Intelligence in Education, 25*(4), 520–542. https://doi.org/10.1007/s40593-015-0064-x

*Miller-Cotto, D. (2017). *The role of prior knowledge, executive function, and perceived cognitive load on the effectiveness of faded worked examples in geometry*. [Doctoral Dissertation, Temple University]. ProQuest Dissertations Publishing.

*Miller-Cotto, D., & Auxter, A. E. (2019). Testing the ecological validity of faded worked examples in algebra. *Educational Psychology, 41*(2), 191–205. https://doi.org/10.1080/01443410.2019.1646411

*Mwangi, W., & Sweller, J. (1998). Learning to solve compare word problems: The effect of example format and generating self-explanations. *Cognition and Instruction, 16*(2), 173–199. https://doi.org/10.1207/s1532690xci1602_2

National Assessment of Educational Progress. (2020). NAEP - 2019 mathematics assessment. https://www.nationsreportcard.gov/mathematics/supportive_files/2019_infographic.pdf

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* U.S. Department of Education.

*Newman, P. M., & DeCaro, M. S. (2019). Learning by exploring: How much guidance is optimal?. *Learning and Instruction*, *62*, 49–63.

* Nye, B. D., Pavlik, P. I., Jr, Windsor, A., Olney, A. M., Hajeer, M., & Hu, X. (2018). SKOPE-IT (Shareable Knowledge Objects as Portable Intelligent Tutors): Overlaying natural language tutoring on an adaptive learning system for mathematics. *International Journal of STEM Education, 5*(1), 12. https://doi.org/10.1186/s40594-018-0109-4

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing instruction and study to improve student learning. IES Practice Guide. NCER 2007–2004. *National Center for Education Research.*

*Reed, S. K., Corbett, A., Hoffman, B., Wagner, A., & MacLaren, B. (2013). Effect of worked examples and cognitive tutor training on constructing equations. *Instructional Science, 41*(1), 1–24. https://doi.org/10.1007/s11251-012-9205-x

*Reiss, K. M., Heinze, A., Renkl, A., & Große, C. (2008). Reasoning and proof in geometry: Effects of a learning environment based on heuristic worked-out examples. *ZDM*, *40*(3), 455-467. https://doi.org/10.1007/S11858-008-0105-0

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23*(1), 90–108. https://doi.org/10.1006/ceps.1997.0959

Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction, 12*(5), 529–556. https://doi.org/10.1016/S0959-4752(01)00030-5

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*(1), 1–37. https://doi.org/10.1111/cogs.12086

Renkl, A., & Atkinson, R. K. (2010). Learning from worked-out examples and problem solving. In J. L. Plass, R. Moreno, & R. Brünken (Eds.)*, Cognitive load theory* (pp. 91–108). Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.007

*Retnowati, E., Ayres, P., & Sweller, J. (2010) Worked example effects in individual and group work settings. *Educational Psychology, 30*(3), 349–367. https://doi.org/10.1080/01443411003659960

*Retnowati, E., Ayres, P., & Sweller, J. (2017). Can collaborative learning improve the effectiveness of worked examples in learning mathematics? *Journal of Educational Psychology*, *109*(5), 666–679. https://doi.org/10.1037/edu0000167

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology, 99*(3), 561–574. https://doi.org/10.1037/0022-0663.99.3.561

Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM Mathematics Education, 49*(4), 599–611. https://doi.org/10.1007/s11858-017-0834-z

Sadler, P. M., & Tai, R. H. (2007). The two high-school pillars supporting college science. *Science, 317*(5837), 457–458. https://doi.org/10.1126/SCIENCE.1144214

*Salden, R. J., Aleven, V., Schwonke, R., & Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. *Instructional Science, 38*(3), 289-307. https://doi.org/10.1007/s11251-009-9107-8

*Schwonke, R., Wittwer, J., Aleven, V., Salden, R. J. C. M., Krieg, C., & Renkl, A. (2007). Can tutored problem solving benefit from faded worked-out examples. In *Proceedings of EuroCogSci*, (Vol. 7, pp. 59–64).

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior, 25*(2), 258–266. https://doi.org/10.1016/j.chb.2008.12.011

Shapka, J. D., Domene, J. F., & Keating, D. P. (2006). Trajectories of career aspirations through adolescence and young adulthood: Early math achievement as a critical filter. *Educational Research and Evaluation, 12*(4), 347–358. https://doi.org/10.1080/13803610600765752

*Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology, 40*, 41–54. https://doi.org/10.1016/j.cedpsych.2014.05.005

Star, J. R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. *Journal of Experimental Child Psychology, 102*(4), 408–426. https://doi.org/10.1016/j.jecp.2008.11.004

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (2006). The worked example effect and human cognition. *Learning and Instruction, 16*(2), 165–169. https://doi.org/10.1016/j.learninstruc.2006.02.005

Sweller, J. (2012). Human cognitive architecture: Why some instructional procedures work and others do not. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, & J. Sweller (Eds.), *APA educational psychology handbook, Vol. 1. Theories, constructs, and critical issues* (pp. 295–325). American Psychological Association. https://doi.org/10.1037/13273-011

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*(1), 59–89. https://doi.org/10.1207/s1532690xci0201_3

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

*Ter Vrugte, J., Jong, T., Vandercruysse, S., Wouters, P., Oostendorp, H., & Elen, J. (2017). Computer game-based mathematics education: Embedded faded worked examples facilitate knowledge acquisition. *Learning and Instruction, 50,* 44–53. https://doi.org/10.1016/j.learninstruc.2016.11.007

Trafton, J. G., & Reiser, B. J. (1993). Studying examples and solving problems: Contributions to skill acquisition [Conference session]. In *Proceedings of the 15th conference of the Cognitive Science Society*, pp. 1017–1022.

*van Harsel, M., Hoogerheide, V., Verkoeijen, P., & van Gog, T. (2019). Effects of different sequences of examples and problems on motivation and learning. *Contemporary Educational Psychology, 58*, 260-275. https://doi.org/10.1016/j.cedpsych.2019.03.005

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology, 47*(1), 513–539.

*van Loon-Hillen, N., van Gog, T., & Brand-Gruwel, S. (2012) Effects of worked examples in a primary school mathematics curriculum, *Interactive Learning Environments, 20*(1), 89–99. https://doi.org/10.1080/10494821003755510

Wallace, B.C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T., A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. Proceedings of the 2nd ACM SIGHT International Health Informatics Symposium (IHI), 819–824.

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*(1), 1–39. https://doi.org/10.1207/s1532690xci0701_1

Zhu, X., & Simon, H. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*(3), 137–166. https://doi.org/10.1207/s1532690xci0403_1

*Zimmerman, W. A. (2015). *Impact of instructional materials eliciting low and high cognitive load on self-efficacy and demonstrated knowledge*. [Doctoral Dissertation, The Pennsylvania State University].

*Zormelo, B. (2018). *Investigating the effect of using worked-out examples teaching approach when teaching properties of special quadrilaterals in grade 10 geometry*. [Unpublished masters thesis]. University of South Africa.