Educational Psychology Review
https://doi.org/10.1007/s10648-021-09610-z

META-ANALYSIS

Multimedia Effect in Problem Solving: A Meta-Analysis

Liru Hu[1] · Gaowei Chen[1] · Pengfei Li[2] · Jing Huang[3]

Accepted: 22 February 2021/Published online: 13 March 2021

Abstract
Pictures are commonly used to represent problems. However, it is unclear how the addition of pictures affects students' problem-solving performance. The multimedia effect in problem solving describes the phenomenon whereby an individual's problem-solving performance is enhanced when equivalent pictures are added to illustrate or replace part of the problem text. Using meta-analytic techniques, this study sought to determine the overall size of the multimedia effect in problem solving and the possible boundary conditions (k = 51; N = 38,987; Range n = 10 – 31,842; Median n = 63). The results showed a significant small-to-medium multimedia effect size on response accuracy (Hedges's g = 0.32) and a significant medium-to-large multimedia effect size on students' response certainty (Hedges's g = 0.74), but no significant multimedia effect on response time. The results for the effects of decorative pictures were not sufficient for a reliable interpretation. Representational (Hedges's g = 0.24) and organizational (Hedges's g = 0.52) pictures had a significant and positive impact on response accuracy, but informational or multiple pictures across studies did not have a significant aggregate effect on an individual's response accuracy. These findings suggest that the multimedia effect in problem solving is diverse and limited by multiple boundary conditions. Further primary studies are needed to further investigate the multimedia effect in problem solving.

Keywords Multimedia effect in problem solving · Response accuracy · Response time · Metacognition · Meta-analysis

Pictures are commonly used in various problem-solving contexts, including formal testing, usual educational exercises, such as classroom practice and homework, and problem-based learning (Greiff et al., 2013; Martin et al., 2016; OECD 2007; Wu et al., 2015). These pictures

✉ Liru Hu
liruhu@connect.hku.hk

[1] Faculty of Education, The University of Hong Kong, 999077 Pok Fu Lam, Hong Kong

[2] School of Education, Shaanxi Normal University, Xi'an 710004, China

[3] Department of Psychology, The Education University of Hong Kong, 999077 Pok Fu Lam, Hong Kong

include pictorial forms such as diagrams, photos, animations, and videos. The way in which problems are represented affects individual problem-solving performance (Bodner and Domin, 2000; Elia et al., 2007; Gagatsis and Elia, 2004). In this meta-analytic review, the *multimedia effect in problem solving* is defined as the phenomenon whereby an individual's problem-solving performance is enhanced when equivalent pictures are added to illustrate or replace part of the problem text. The multimedia effect in problem solving is assumed to exist in various types of problems across different domains, such as structured multiple-choice problems (e.g., Lindner et al., 2017a, 2017b; Saß et al., 2012), word problems (e.g., Dewolf et al., 2015; Hoogland et al., 2018a), insight problems (Schwert, 2007), and analogy problems (e.g., Beveridge and Parkins, 1987). It includes problem solving for both assessment and learning purposes.

Compared with the multimedia effect in learning or the multimedia principle (Mayer, 2001, 2009, 2014), which has been the subject of extensive research (Butcher, 2014; Rop, 2017; Schüler et al., 2019), the multimedia effect in problem solving remains a relatively new topic of research (Lindner et al., 2018). Fairly recently, a growing number of studies have investigated the extent to which pictures affect an individual's problem-solving performance, often through randomized experimental designs or qualitative analysis of individuals' solution strategies. These studies have involved various subjects such as mathematics (e.g., Agathangelou et al., 2008; Berends and van Lieshout, 2009; Dewolf et al., 2017; Elia et al., 2007; Elia and Philippou, 2004; Hoogland et al., 2018a; Ögren et al., 2017; Zahner and Corter, 2010), science (e.g., Hao, 2010; Hartmann and Leutner, 2013; Lindner et al., 2017a, 2018; Lindner et al., 2017b; Maries, 2013; Whitley et al., 2006), and language (e.g., Dindar et al., 2013; Garrett, 2008; Ginther, 2001). However, due to inconsistent findings, it remains unclear how the addition of pictures affects students' problem-solving performance (Wu et al., 2015). To date, no comprehensive meta-analytic review has been conducted to determine the aggregate multimedia effect in problem solving and its relevant boundary conditions. This study aims to fill this gap and provide insights into the design of text–picture problems.

Theoretical Background

Different isomorphic external representations of a problem may generate dramatically different cognitive processes and problem-solving behaviors, which is referred to as the representational effect (Chuah et al., 2000; Zhang and Norman, 1994, 1995). A problem can be represented in a pictorial (i.e., picture) or descriptive (i.e., word) format (Schnotz and Bannert, 2003). External representations of a problem refer to the presentation of the problem to the problem solver, while internal representations describe how the problem solver constructs the problem in his or her mind, also known as a mental model (Johnson-Laird, 2005). Students can also externalize their internal representations through drawing, writing or speaking in the problem-solving process.

There are currently no specific theories to explain the role of pictures in problem solving (Lindner et al., 2017b). The problem-solving process involves mental model construction similar to the learning context. Thus, many scholars have relied on intensively studied multimedia learning theories to explain the benefits of pictures in problem solving (Jarodzka et al., 2015; Lindner et al., 2016). The most frequently cited multimedia learning theories include the cognitive theory of multimedia learning (CTML) (Mayer, 2001, 2009, 2014), the cognitive–affective theory of learning with multimedia (CATLM) (Moreno, 2006, 2009;

Moreno and Mayer, 2007), the integrated model of text and picture comprehension (Schnotz and Bannert, 2003), and cognitive load theory (CLT) (Sweller, 1988; Sweller, 1994). These theories basically assume that individuals construct internal representations of an object by encoding features observed in an environment (Newell, 1990). They attribute the multimedia effect in learning to the advantage of dual-channel processing (Baddeley, 1992; Baddeley et al., 2011; Paivio, 1986), that is, the encoding of texts and pictures through two symbolic systems: auditory/verbal and visual/pictorial. Each system is limited in terms of working memory capacity, but their combined capacity is additive (Chandler and Sweller, 1991) (see Fig. 1). Individuals can access greater processing capacity when materials are presented through the two symbolic systems.

Students are also more likely to construct better mental models through dual-channel processing, as visual–spatial and verbal representations complement and constrain each other, and thus deepen students' knowledge construction (Ainsworth, 2006; Mayer, 2014). Mayer (2009, 2014) further differentiated five cognitive processes underlying multimedia learning: selecting relevant words into verbal working memory; selecting relevant pictures into visual working memory; organizing selected words into a verbal model; organizing selected pictures into a pictorial model; and integrating the verbal model, the pictorial model and activated prior knowledge. However, the added value of integration is conditional and requires additional efforts. Therefore, there are boundary conditions for the multimedia effect in learning (Schnotz and Bannert, 2003; Mayer, 2014; Schüler et al., 2019). Although these multimedia learning theories can help explain the multimedia effect in problem solving to some extent, very few scholars have followed multimedia learning principles when designing text–picture problems (Hoogland et al., 2018a).

There are a number of stepwise models that describe the problem-solving process across different domains (e.g., Mayer, 1987; Nathan et al., 1992; Reusser, 1996; Verschaffel et al., 2000; Winter et al., 2006; Zahner and Corter, 2010). The problem-solving process includes not only an internalization process (comprehension phase) but also an externalization process (solution phase), which is different from the learning process (see Fig. 1). Thus, pictures in problem solving have two functions. One function is similar to that in the learning context, that is, serving as mental scaffolds to interpret question items (Lindner et al., 2017a, 2017b) and make the most out of our working memory capacity (Saß et al., 2012). The other function is to facilitate the extraction of information in the answering process and thus to offload problem solving (Lindner et al., 2017b; Zhao et al., 2014). Pictures are easier to remember and recall than words (Nickerson, 1965; Shepard, 1967) and provide easier access to semantic memory (Paivio, 1986). Thus, pictures can serve as prominent recall cues and facilitate the solution phase.

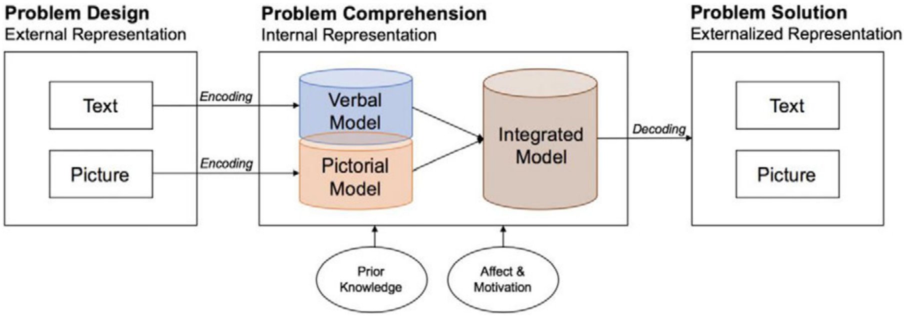


Fig. 1 Explanation of the multimedia effect in problem solving based on multimedia learning theories

Instead of assuming that cognition occurs solely in the human brain, the theories of embodied and distributed cognition posit that human cognition is distributed throughout the human brain, the human body and the external environment (Hardy-Vallée and Payette, 2009; Hollan et al., 2000). Such theoretical approaches have been adopted to explain human cognitive processing in interaction with multiple external representations (Pande and Chandrasekharan, 2017). External representations can be a way to offload an individual's cognitive load by providing ready-to-use representations (Zhang and Patel, 2006). Individuals do not necessarily construct internal representations to mediate actions, but they can act directly on information extracted from external representations (Gibson, 1966, 1979). Therefore, internal and external representations are not differentiated but are instead indispensable and distributed across the human mind and cognitive artifacts (Zhang, 1997; Zhang and Norman, 1994) (see Fig. 2). Pictures are advantageous for expressing visual–spatial information; they are more closely aligned with the mental models of problem solvers (Schnotz, 2014). Thus, task-relevant pictures in problems are effective in scaffolding a student's problem-solving process. However, the benefits of additional pictorial representations depend on whether students can make full use of these external representations and generate their own by adjusting those provided.

Measuring the Multimedia Effect in Problem Solving

Two major quantitative cognitive outcomes (i.e., response accuracy and response time) are used to measure the multimedia effect in problem solving. Response accuracy measures how close an answer is to the correct answer or the proportion of correct answers in a large problem set. Response time measures the duration of perceptual, cognitive, and psychometric activities to generate a solution (Goldhammer et al., 2014). In addition to studies on quantitative

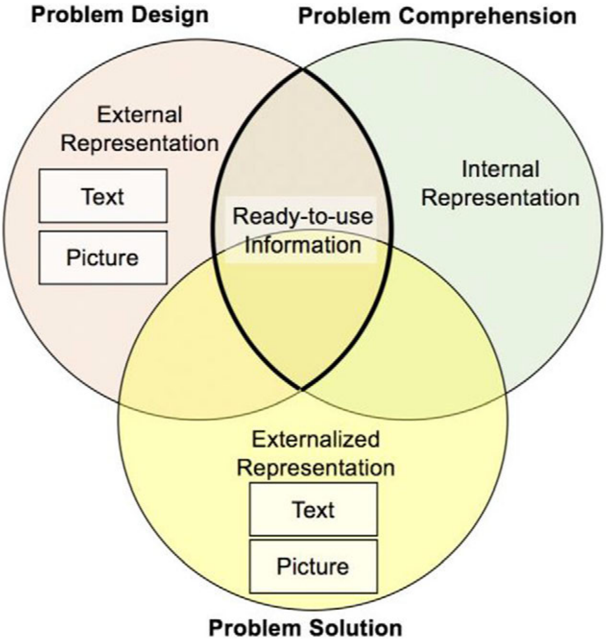


Fig. 2 Explanation of the multimedia effect in problem solving through the embodied and distributed cognition approaches

cognitive outcomes, some studies have explored the multimedia effect on an individual's affective experience or metacognition, including pleasure and engagement in problem solving, preference for pictures, perceived task difficulty, and response certainty (i.e., students' level of confidence in their response) (e.g., Dewolf et al., 2015; Lindner et al., 2018).

Multimedia Effect on Response Accuracy

Response accuracy has been widely used as an indicator of an individual's problem-solving performance and is typically measured in relation to the difficulty of an item (e.g., Lindner et al., 2017b). Response accuracy and item difficulty are distinct but strongly correlated. Thus, the multimedia effect on response accuracy in this study included the multimedia effect on individual accuracy in problem solving and item difficulty levels.

Measurements of the multimedia effect on response accuracy have yielded inconsistent results. Some findings have suggested that representational pictures make items easier to understand, even without providing solution-relevant information (Lindner et al., 2016; Lindner et al., 2017b; Saß et al., 2012). This phenomenon is referred to as the multimedia effect in testing (Lindner et al., 2018; Lindner et al., 2017a). For example, Lindner et al., (2017b) added pictures to illustrate the stems of multiple-choice questions and found that the test performance of all students improved. Hoogland et al., (2018a) added pictures to math word problems in an integrated format and also observed improvement in student performance.

However, several studies have found no multimedia effect on response accuracy (Agathangelou et al., 2008; Dewolf et al., 2014; Dewolf et al., 2015; Dewolf et al., 2017). For example, the study conducted by Dewolf (2014) on the use of visual aids to support students in solving non-standard math word problems yielded non-significant results. Some studies have even detected a reverse multimedia effect on response accuracy (Berends and van Lieshout, 2009; Elia et al., 2007; Gagatsis and Elia, 2004).

Multimedia Effect on Response Time

Response time, also called "time on task," is a process measure of students' processing speed, regardless of accuracy (Goldhammer et al., 2014; Scherer et al., 2015). Researchers have focused on whether pictures can boost students' problem-solving fluency. However, measuring the multimedia effect through response time has yielded inconsistent results. Some researchers have reported a reduction in response time when students are asked to solve text–picture problems rather than text-only problems (Lindner et al., 2017b; Saß et al., 2012), while others have found no such effect (e.g., Berends and van Lieshout, 2009; Lindner et al., 2017a).

On the one hand, adding pictures to the problem text may complicate the problem and increase the response time by providing more information to process. Attention distraction is a general effect of pictures (Harp and Mayer, 1998; Wiley et al., 2014). Indeed, eye-tracking research has shown that pictures compete for students' attention in problem solving to varying degrees (Dewolf et al., 2015; Lindner et al., 2017a). Studies on seductive details (i.e., eye-catching but irrelevant information) in multimedia learning have also shown that students spend more time learning when there are seductive details in instructional materials (Park et al., 2011). On the other hand, some eye-tracking studies have shown that more information does not necessarily extend the response time. Additional pictures in problems may capture very little student attention (Dewolf et al., 2015). In addition, more time spent on processing pictures may be compensated by less time spent on processing texts (Lindner et al., 2017a).

Meanwhile, adding pictures may increase students' problem-solving fluency according to dual processing theory, which distinguishes two mental processes: automatic and controlled processes (Schneider and Chein, 2003; Shiffrin and Schneider, 1977). Pictures physically resemble an object by visually representing many of its specific and idiosyncratic features. This representational process is more akin to low-level construal. By contrast, words highlight the categorical invariant characteristics of an object, which is more akin to high-level construal. Thus, pictures tend to be processed automatically and rapidly, requiring little mental effort, whereas words are processed through a much more controlled and conscientious process that requires higher-level cognitive operations. As such, adding pictures to illustrate or replace part of the problem text may speed up students' mental model construction. According to the picture superiority effect (Nickerson, 1965; Shepard, 1967), pictures are also better recall cues in information retrieval and thus accelerate solution generation. In addition, pictures can facilitate the answering process through analog structure mapping (e.g., Beveridge and Parkins, 1987; Gentner et al., 2003; Schnotz and Bannert, 2003).

Multimedia Effect on Metacognition and Affective Experience

In contrast to the diverse findings concerning the multimedia effect on response accuracy and response time, existing studies have consistently found that pictures in problems generate higher metacognitive judgments and positive affective experiences.

Students have been found to perceive lower item difficulty (Garcia-Retamero and Holffrage, 2013) and higher response certainty when solving text–picture problems than text-only problems (Lindner et al., 2018; Dewolf et al., 2015; Ögren et al., 2017). According to dual processing theory (Schneider and Chein, 2003; Shiffrin and Schneider, 1977) and the picture superiority effect (Nickerson, 1965; Shepard, 1967), pictures tend to boost perceptual fluency more than words and are easier to remember and retrieve from memory. Moreover, increased perceptual and retrieval fluency tend to induce higher metacognitive judgments (Bjork et al., 2013). Therefore, pictures are more likely to make the item subjectively easier to understand compared with text, boosting students' confidence in problem solving.

Research has also found that pictures in problems tend to lead to positive affective experiences. For example, students are more likely to report more positive attitudes and higher test-taking pleasure when pictures are embedded in problem texts (Hao, 2010; Lindner et al., 2016). Students also tend to show greater situational interest (Ögren et al., 2017) and a higher level of arousal when presented with engaging pictorial representations.

The multimedia effect on metacognition and affective experience may also influence the multimedia effect on response accuracy and response time by influencing students' cognitive engagement and behavior in problem solving. Pictures may reduce individuals' rapid-guessing behavior in testing (Lindner et al., 2017b; Wise et al., 2009), increase their mental effort in problem solving (Ögren et al., 2017), persuade students to believe an accompanying statement (Ögren et al., 2017) or possibly slow down the loss of engagement in problem solving (Lindner et al., 2017b).

Moderators of the Multimedia Effect in Problem Solving

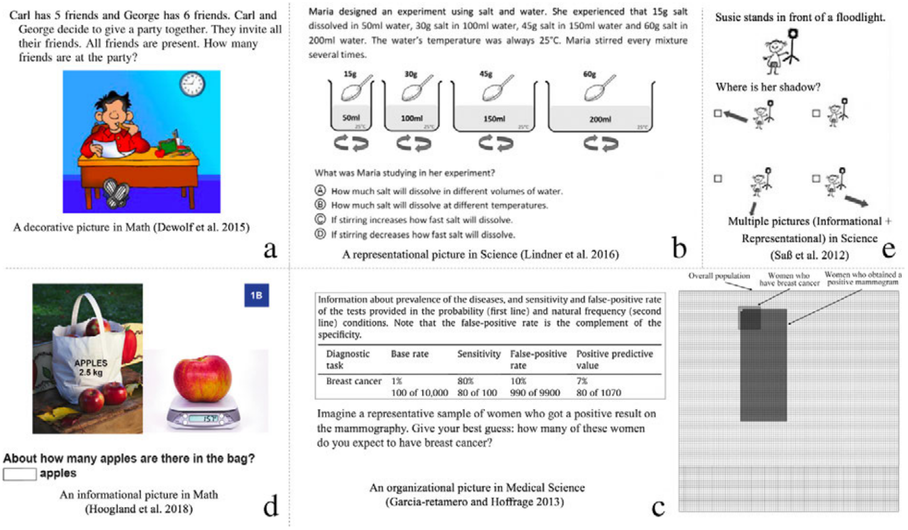
Function Types of Pictures

The multimedia effect in problem solving is conditional upon the types of pictures included in problem-solving tasks. Studies have compared how pictures of various function types affect the multimedia effect in problem solving (e.g., Agathangelou et al., 2008; Berends and van

Lieshout, 2009; Dewolf et al., 2015; Elia et al., 2007; Saß et al., 2012; Saß et al., 2017). However, the function types of pictures have been mainly investigated in isolation across studies.

Specifically, pictures in problems have been categorized according to four functions: decorative, representational, organizational, and informational (Berends and van Lieshout, 2009; Carney and Levin, 2002; Levin, 1981; Saß et al., 2017) (see Fig. 3). Decorative pictures have little or no relation to the text and are usually aimed at triggering a positive mood or creating a learning atmosphere for the problem solver, for example, in the study by Dewolf et al., (2015), a cartoon of a student pondering a word problem was added to a text-based test (see Fig. 3a). Representational pictures are concrete illustrations of information conveyed in the text. They are closely related to the context of the problem and serve to scaffold an individual’s understanding or to strengthen his or her engagement in the task. For example, Lindner et al. (2016) visualized the components of four types of solutions to a chemical solution problem using schematic diagrams (see Fig. 3b). Organizational pictures present the information structures in the text through a visual–spatial approach. For example, in the study by Garcia-Retamero and Hoffrage (2013), in a diagnostic inference task, a square map reflected the natural frequency of breast cancer, including the overall number of women at risk, the number of women who had breast cancer and the number of women who had received a positive mammogram (see Fig. 3c). Informational pictures contain details needed to solve the problem, such as diagrams and charts. For example, in the study by Hoogland et al. (2018a), texts and pictures were integrated in such a way that individuals would be unable to solve a problem if the pictures were removed (see Fig. 3d).

Some problems involve several pictures that serve different functions. For example, in the study by Saß et al. (2012), in a multiple-choice science question, a picture of a girl with a floodlight in the item stem was representational, while a girl with a floodlight and her possible shadow was informational in the answer choices (see Fig. 3e). Specific pictures are usually categorized according to their primary function.



Carl has 5 friends and George has 6 friends. Carl and George decide to give a party together. They invite all their friends. All friends are present. How many friends are at the party?

A decorative picture in Math (Dewolf et al. 2015)

Maria designed an experiment using salt and water. She experienced that 15g salt dissolved in 50ml water, 30g salt in 100ml water, 45g salt in 150ml water and 60g salt in 200ml water. The water's temperature was always 25°C. Maria stirred every mixture several times.

What was Maria studying in her experiment?

Ⓐ How much salt will dissolve in different volumes of water.
Ⓑ How much salt will dissolve at different temperatures.
Ⓒ If stirring increases how fast salt will dissolve.
Ⓓ If stirring decreases how fast salt will dissolve.

A representational picture in Science (Lindner et al. 2016)

Susie stands in front of a floodlight.

Where is her shadow?

□ □
□ □

Multiple pictures (Informational + Representational) in Science (Saß et al. 2012)

About how many apples are there in the bag?

□ apples

An informational picture in Math (Hoogland et al. 2018)

Information about prevalence of the diseases, and sensitivity and false-positive rate of the tests provided in the probability (first line) and natural frequency (second line) conditions. Note that the false-positive rate is the complement of the specificity.

Diagnostic task	Base rate	Sensitivity	False-positive rate	Positive predictive value
Breast cancer	1% 100 of 10,000	80% 80 of 100	10% 990 of 9900	7% 80 of 1070

Imagine a representative sample of women who got a positive result on the mammography. Give your best guess: how many of these women do you expect to have breast cancer?

An organizational picture in Medical Science (Garcia-retamero and Hoffrage 2013)

Fig. 3 Overview of the different types of pictures in problems. **a** Decorative picture; **b** Representational picture; **c** Organizational picture; **d** Informational picture; **e** Multiple pictures

Researchers have investigated how different pictorial functions affect individuals' problem-solving performance (Agathangelou et al., 2008; Carney and Levin, 2002; Elia et al., 2007; Elia and Philippou, 2004; Gagatsis and Elia, 2004; Lindner et al., 2016, 2018). Pictures that contain task-relevant information are more likely to reduce item difficulty and improve problem-solving performance, while adding decorative pictures to the text tends to neither enhance nor harm problem-solving performance (Agathangelou et al., 2008; Carney and Levin, 2002; Elia et al., 2007; Lindner et al., 2018).

Other Moderators

In addition, the way in which texts and pictures are combined is important for problem-solving performance. Functionally related texts and pictures that are placed separately may lead to a split-attention effect and extraneous cognitive processing (Berends and van Lieshout, 2009; Crisp and Sweiry, 2006; Ögren et al., 2017). Therefore, an individual's problem-solving performance may worsen if texts and pictures providing complementary information are separated in space or time (Saß and Schütte, 2016; Saß et al., 2017). An individual's test performance is also likely to be affected when irrelevant content is provided (Jarodzka et al., 2015).

Researchers have also considered the moderating effect of problem complexity (Hoogland et al., 2018b; Lindner et al., 2016; Solano-Flores et al., 2016; Zahner and Corter, 2010). A wordy problem tends to overload students' cognitive capacity, as it requires better reading comprehension, which may adversely affect their problem-solving performance. Pictures can make the spatial structure of a wordy problem explicit and thus easier to understand (Boonen et al., 2014; Fuchs et al., 2015; Hoogland et al., 2018b), although some studies have not detected such an effect, which may be due to the low variance in the problem's wordiness (e.g., Lindner et al., 2016).

The multimedia effect in problem solving may also differ across domains. External representations in different domains vary considerably (Pande and Chandrasekharan, 2017). For example, representations in chemistry can include the periodic table, chemical equations, molecular diagrams, energy graphs, simulations, and so on, (Johnstone, 1982), while in mathematics, they can include numbers, algebraic equations, geometries, and metric spaces (Lesh et al., 1987). Although this study did not focus on a specific domain due to the limited number of primary studies, the domain was considered a moderator in the following analysis.

Research Hypotheses

This meta-analytic review aimed to bring together current quantitative research on the multimedia effect in problem solving using the three outcome measures, and to identify relevant moderators. In terms of study outcomes related to metacognition and affective experience, the statistics available on the measurement of affective experience are not sufficient to conduct a meta-analysis. Therefore, this study only focused on the multimedia effect on metacognition, in particular response certainty.

Two hypotheses are proposed based on the literature:

Hypothesis 1. The multimedia effect on response accuracy: Individuals have greater response accuracy when solving text–picture problems than when solving text-only problems.

Hypothesis 2. The multimedia effect on metacognition: Individuals have greater response certainty when solving text–picture problems than when solving text-only problems.

Regarding the multimedia effect on response time, that is, whether additional pictures speed up students' problem solving, there are no clear preferred expectations in either the empirical or theoretical accounts. Therefore, no hypothesis is proposed on this effect, but the effect was carefully examined based on the included studies.

Method

Data Collection and Reduction

This study followed Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) to identify relevant studies (Moher et al., 2009). Initial searches adopted the following search terms and their variants to limit the titles or subjects: ("multiple representations" OR "text–picture" OR "multiple external representations" OR "image-rich" OR "illustration" OR "picture" OR "visual representation" OR "pictorial representation" OR "depictive representation") AND ("problem solving" OR "problem-solving" OR test* OR "multimedia effect in testing" OR "multimedia testing" OR "multimedia problem solving"). The initial screening found two thousand, nine hundred and eighty-four articles (2984) in the ProQuest Dissertations and Theses A&I, Eric, LISA, PsycARTICLES, PsycINFO, PsycTESTS, ScienceDirect, Web of Science, and China National Knowledge Infrastructure databases. There was no start date limit, but the cut-off date was set to before August 8, 2018.

To be included in the final meta-analysis, studies had to meet the following inclusion and exclusion criteria:

(a) The study should use an experimental or quasi-experimental design, with a control group in the text-only condition and at least one experimental group in the text–picture condition;
(b) The problems in the text-only and text–picture conditions should be equivalent, that is, adding pictures or modifying texts should not affect solution-relevant information;
(c) At least one of the following target outcomes should be measured quantitatively: response accuracy, response time or affective experience;
(d) The quantitative results should include the effect size for any specific measure or enough data to compute the effect size;
(e) When multiple publications are based on the same data, only one publication should be included;
and
(f) The study should be reported in English or Chinese.

In addition to these search and screening procedures, Google Scholar was searched using the main keywords to identify conference abstracts, government documents, and other online materials. Snowballing from the reference lists of the primary studies helped identify other relevant research and new keywords. This strategy was essential because this meta-analysis topic is relatively new and relevant studies are distributed across different disciplines, so they tend to use inconsistent terminologies to describe similar phenomena.

The entire screening procedure is detailed in the PRISMA flow diagram (see Fig. 4). The initial coder excluded “easy rejects” through title screening. Another coder was then trained and filtered the remaining papers according to the inclusion and exclusion criteria independently. Overall inter-rater reliability (Cohen’s kappa) for abstract and full-text screening was high ($\kappa = 0.86$). All disagreements were discussed and consensus was reached. The first coder then finished the snowball search and contacted a few key researchers who published many papers in the field (e.g., Lindner, Hoogland, Van Dooren and Dewolf) to check whether there were any unpublished studies or important publications missing from the final literature list.

Coding Procedure and Study Characteristics

Codes were developed to characterize the studies included in the review list (see Table 1). Two researchers conducted the coding work independently. The inter-rater reliability of the main moderators in terms of initial coding was high ($\kappa > 0.80$). All disagreements were resolved by seeking consensus.

Computation of Effect Sizes and Analysis

Hedges’s *g* was adopted to measure the effect sizes. R (a software environment for statistical computing and graphics) as well as the Comprehensive Meta-Analysis (CMA, version 2) software were used to conduct the meta-analyses in this study. For between-groups designs, the data formats for calculating the effect sizes included “means, SDs, and sample sizes,” “Cohen’s *d* and variance,” “events and sample size in each group,” “means, sample sizes, and

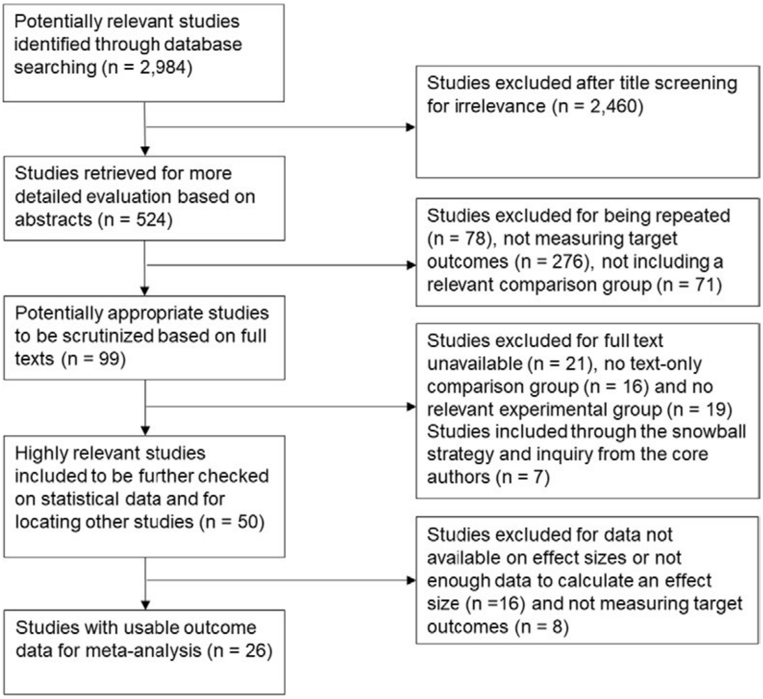

Potentially relevant studies identified through database searching (n = 2,984)
Studies excluded after title screening for irrelevance (n = 2,460)
Studies retrieved for more detailed evaluation based on abstracts (n = 524)
Studies excluded for being repeated (n = 78), not measuring target outcomes (n = 276), not including a relevant comparison group (n = 71)
Potentially appropriate studies to be scrutinized based on full texts (n = 99)
Studies excluded for full text unavailable (n = 21), no text-only comparison group (n = 16) and no relevant experimental group (n = 19) Studies included through the snowball strategy and inquiry from the core authors (n = 7)
Highly relevant studies included to be further checked on statistical data and for locating other studies (n = 50)
Studies excluded for data not available on effect sizes or not enough data to calculate an effect size (n =16) and not measuring target outcomes (n = 8)
Studies with usable outcome data for meta-analysis (n = 26)

Fig. 4 PRISMA flow diagram of article selection for problem solving with pictorial representations

Table 1 Codes for study characteristics

Types	Codes	Descriptions
Basic information	Authors	
	Year of publication	–
	Publication type	This includes *journal article*, *conference article*, *thesis*, *unpublished article* and *book chapter*.–
	Institution location	This refers to the country/region where the institution of the first author is located. Due to the limited number of included studies, this code was further divided into *US*, *Europe* and *other*.
Statistical information	This includes all statistical data available in the articles (e.g., effective sample sizes, mean values, standard deviations, standard errors, p values, different model parameters) to compute Hedges's g, z values, p values, and 95% confidence intervals. The effective sample size refers to the effective number of units of analysis to calculate an effect size and may be inconsistent with the number of participants.	
Study design	Outcome	This includes *response accuracy*, *response time*, and *response certainty*.
	Function type	This refers to the type of function that the pictures serve in the problem. It includes *decorative*, *representational*, *organizational*, *informational,* and *multiple*.
	Set size	This describes the number of problems in the experiment. Most of the included studies adopted paired text-only problems and text–picture problems. The set size here only refers to the number of problems in one group.
	Problem difficulty	This differentiates two levels of problem difficulty: *easy* and *difficult*. It was coded based on explicit clues in the articles or on the control group's response accuracy if there were no such clues. Problems with an average accuracy percentage less than 40% were coded as *difficult*, and *easy* otherwise. Codes for the same item(s) were held constant in each study.
	Domain	This includes *Mathematics*, *Science*, *Language,* and *Medicine*.
	Answer format	This includes multiple-choice, fill-in-the-blank, true or false, matching and structured open responses. We further divided this category into two sub-categories: *multiple-choice* in which the correct answer was provided and *structured open response* in which a single answer was correct and students needed to find it.
	Test environment	This refers to the location of the experiment (*regular classroom*, *lab*, *online* or *other*) and the medium in which the experiment was delivered (*paper and pencil* or *computer-based*).
	Number of participants	The actual number of participants involved in the study. This does not include those who were excluded by the authors.
	Education level	This refers to the education level of the participants, including *primary*, *university,* and *others.*
	Experimental design	This includes *between-groups design*, *within-groups design,* and *others*.
	Control conditions	This refers to whether a study controlled some variable(s) or not.

p-values" and "sample sizes and *p*-values." The data formats for within-group designs included "means, sample sizes, and paired *t*-values," "mean difference (paired) and standard error" and "sample sizes and *p*-values from paired *t* tests."

To control for interdependencies between effect sizes, the effect sizes for the same target outcome in a single study were combined before the studies were pooled if the participants involved in these effect sizes overlapped. However, for the moderator analysis, different effect sizes in one study were retained if they were related to different types of pictures or to item(s) with different levels of difficulty. Influential and outlying studies were also identified using the Baujat plot (Baujat et al., 2002) before computing the summary effect sizes for different outcomes. A variety of outlier and influential case diagnostics can be computed when conducting a meta-analysis (Viechtbauer and Cheung, 2010). A Baujat plot visually identifies sources of heterogeneity in meta-analytic data by plotting each study's contribution to overall heterogeneity against each study's contribution to the combined effect size (Baujat et al., 2002).

Based on previous studies, multiple variables, such as picture functions, problem difficulty, and problem domain, were predicted to influence the multimedia effect in problem solving. Thus, random effects and mixed effects models were adopted to conduct the main meta-analysis and moderation analysis using the *rma* function in the *metafor* package (Viechtbauer and Cheung, 2010). Regarding publication bias, funnel plots (Light and Pillemer, 1984; Sterne and Egger, 2001) were used for all of the outcomes (Rothstein et al., 2005). Furthermore, the *ranktest* and *regtest* functions were used to detect funnel plot asymmetry. For significant observations, Rosenthal's fail-safe N test was conducted using the *fsn* function in R. The fail-safe number estimates any possible missing publication with an average *z*-value of zero that could make the combined effect size non-significant (Rosenthal, 1979).

As a general rule, all meta-analysis results should be interpreted with caution because, by necessity, they only select quantitative evidence (Greco et al., 2013). Furthermore, the combined effect sizes may not be robust enough if the available studies are too few or too different. It is thus essential to examine the heterogeneity of the observed effect sizes in a meta-analysis. Different study characteristics may contribute to differences between studies that could not be explained by subject-level sampling errors (Lipsey and Wilson, 2001). Therefore, if there is heterogeneity, moderator analysis is usually performed in a meta-analysis to help explain the heterogeneity of the observed effect sizes. This study reported Q and I^2 statistics to test for heterogeneity of variance. If Q exceeds a critical value ($p < .05$), the average effect size is interpreted as significantly heterogeneous. I^2 estimates the variance of aggregate effect sizes that is due to heterogeneity rather than sampling errors. I^2 at around 25%, 50%, and 75% is interpreted as indicating low, medium, and high heterogeneity, respectively (Higgins and Thompson, 2002).

Results

Sample Analysis

There were 51 pairwise comparisons from 26 articles (see Table 2): 23 journal articles, two theses (Garrett, 2008; Maries, 2013), and one research report (Ginther, 2001). The very low number of original articles limited the reliability of aggregate effect sizes in this meta-analysis. All of the publications included in the final meta-analysis were in English. Most of the authors were from Europe (54%) and the USA (38%). The publication dates ranged from 1983 to 2018. Over 60%

Table 2 Overview of included studies[a]

Study name	Outcome	Domain	Design[b]	Education level	Problem difficulty	Environment[c]	Location	Answer format[d]	Set size	Function type	Effective sample size[e]	Number of participants	Have control variable	Hedges's g[f]
Berends and van Lieshout, 2009	Time	Math	W	Primary	Easy	PP	Euro	SOR	24	Decorative	130	130	Yes	−0.17
Berends and van Lieshout, 2009	Accuracy	Math	W	Primary	Easy	PP	Euro	SOR	24	Decorative	130	130	Yes	0.001
Beveridge and Parkins, 1987	Accuracy	Science	B	Primary	Difficult	PP	Euro	SOR	1	Organizational	120#	145	Yes	0.28
Beveridge and Parkins, 1987	Accuracy	Science	B	University	Easy	PP	Euro	SOR	1	Organizational	91	91	Yes	0.37
Brase, 2008	Accuracy	Medicine	B	University	Difficult	PP	US	SOR	2	Organizational	289	289	No	0.10#
Dewolf et al., 2014	Accuracy	Math	B	Primary	Difficult	PP	Euro	SOR	10	Representational	402	402	Yes	0.09
Dewolf et al., 2014	Accuracy	Math	B	Primary	Difficult	PP	Euro	SOR	8	Representational	233	233	Yes	0.03
Dewolf et al., 2015	Response certainty	Math	B	University	Difficult	Com	Euro	SOR	8	Decorative	20	20	No	1.78
Dewolf et al., 2015	Response certainty	Math	B	University	Difficult	Com	Euro	SOR	8	Representational	20	20	No	0.92
Garcia-Retamero et al., 2010	Accuracy	Medicine	B	Other	Easy	Com	Euro	SOR	8	Organizational	116#	117	No	0.69
Garcia-Retamero and Hoffrage, 2013	Accuracy	Medicine	B	Other	Difficult	PP	Euro	SOR	3	Organizational	156#	162	Yes	1.00
Garcia-Retamero and Hoffrage, 2013	Response certainty	Medicine	B	Other	Difficult	PP	Euro	SOR	3	Organizational	156#	162	Yes	0.41
Garrett, 2008	Accuracy	Language	B	Primary	Easy	PP	US	MC	43	Informational	64	64	Yes	0.08
Gick and Holyoak, 1983	Accuracy	Science	B	Other	Difficult	PP	US	SOR	1	Organizational	51	51	No	−0.43
Gick and Holyoak, 1983	Accuracy	Science	B	Other	Difficult	PP	US	SOR	1	Organizational	189	189	No	0.45
Ginther, 2001	Accuracy	Language	W	Other	Easy	Com	US	SOR	40	Multiple	160	160	No	0.07

Table 2 (continued)

Study name	Outcome	Domain	Design[b]	Education level	Problem difficulty	Environment[c]	Location	Answer format[d]	Set size	Function type	Effective sample size[e]	Number of participants	Have control variable	Hedges's g[f]
Goolkasian, 1996	Accuracy	Math	W	University	Easy	Com	US	MC	16	Informational	28	28	Yes	0.84
Goolkasian, 1996	Accuracy	Math	W	University	Easy	Com	US	MC	48	Informational	10	10	Yes	0.36
Goolkasian, 1996	Time	Math	W	University	Easy	Com	US	MC	16	Informational	28	28	Yes	0.84
Goolkasian, 1996	Time	Math	W	University	Easy	Com	US	MC	48	Informational	10	10	Yes	1.91
Hoogland et al., 2018a	Accuracy	Math	W	Other	Difficult	Com	Euro	SOR	21	Informational	31,842	31,842	Yes	0.09
Lin et al., 2013; Yang and Huang, 2004	Accuracy	Math	W	Primary	Easy	PP	US	MC	16	Informational	681	681	No	−0.72#
Lindner et al., 2016	Item difficulty	Science	W	Primary	Easy	PP	Euro	MC	30	Representational	30#	305	Yes	0.52
Lindner et al., 2017a	Item difficulty	Science	W	Primary	Easy	Com	Euro	MC	24	Representational	24#	62	Yes	0.58
Lindner et al., 2017a	Item time	Science	W	Primary	Easy	Com	Euro	MC	24	Representational	24#	62	Yes	−0.13
Lindner et al., 2017b	Accuracy	Science	W	Primary	Easy	Com	Euro	MC	36	Representational	401	401	No	0.13
Lindner et al., 2017b	Time	Science	W	Primary	Easy	Com	Euro	MC	36	Representational	401	401	No	−0.11
Lindner et al., 2018	Response certainty	Science	B	University	Easy	PP	Euro	MC	33	Representational	63	63	Yes	0.54
Lindner et al., 2018	Accuracy	Science	B	University	Difficult	PP	Euro	MC	16	Representational	63	63	Yes	0.35
Lindner et al., 2018	Accuracy	Science	B	University	Easy	PP	Euro	MC	33	Representational	63	63	Yes	0.75
Maries, 2013	Accuracy	Science	B	University	Easy	PP	US	SOR	5	Representational	120	120	Yes	−0.62#
Ögren et al., 2017	Accuracy	Math	B	University	Easy	Com	Euro	MC	8	Representational	36	36	No	0.21
Ögren et al., 2017	Time	Math	B	University	Easy	Com	Euro	MC	8	Representational	36	36	No	−0.02
Ott et al., 2018	Accuracy	Math	B	University	Difficult	PP	Euro	MC	19	Representational	47	47	Yes	0.32
Ott et al., 2018	Accuracy	Math	B	University	Difficult	PP	Euro	MC	19	Representational	50	50	Yes	1.29
Ott et al., 2018	Accuracy	Math	B	University	Difficult	PP	Euro	MC	19	Representational	49	49	Yes	0.09
Ramjan, 2011	Accuracy	Math	W	University	Easy	PP	Other	SOR	12	Multiple	567	567	No	0.98

Table 2 (continued)

Study name	Outcome	Domain	Design[b]	Education level	Problem difficulty	Environment[c]	Location	Answer format[d]	Set size	Function type	Effective sample size[e]	Number of participants	Have control variable	Hedges's g[f]
Saß et al., 2012	Accuracy	Science	B	Primary	Easy	Com	Euro	MC	22	Representational	158	158	Yes	0.27#
Saß et al., 2012	Accuracy	Science	B	Primary	Easy	Com	Euro	MC	22	Representational	58	58	Yes	0.63
Saß et al., 2012	Accuracy	Science	B	Primary	Easy	Com	Euro	MC	22	Informational	57	57	Yes	0.08
Saß et al., 2012	Accuracy	Science	B	Primary	Easy	Com	Euro	MC	22	Informational	54	54	Yes	0.30
Saß et al., 2012	Accuracy	Science	B	Primary	Easy	Com	Euro	MC	22	Multiple	55	55	Yes	0.62
Saß et al., 2012	Time	Science	B	Primary	Easy	Com	Euro	MC	22	Representational	158	158	Yes	0.16#
Saß et al., 2012	Time	Science	B	Primary	Easy	Com	Euro	MC	22	Representational	58	58	Yes	0.07
Saß et al., 2012	Time	Science	B	Primary	Easy	Com	Euro	MC	22	Informational	57	57	Yes	0.19
Saß et al., 2012	Time	Science	B	Primary	Easy	Com	Euro	MC	22	Informational	54	54	Yes	0.61
Saß et al., 2012	Time	Science	B	Primary	Easy	Com	Euro	MC	22	Multiple	55	55	Yes	0.26
Schwert 2007	Time	Science	W	University	Difficult	PP	US	SOR	12	Informational	144	144	No	−0.19
Solano-Flores et al., 2014	Accuracy	Science	W	Other	Easy	PP	US	MC	27	Representational	728	728	No	0.08
Whitley et al., 2006	Accuracy	Science	B	University	Easy	Com	US	SOR	9	Organizational	31	31	Yes	1.11
Whitley et al., 2006	Time	Science	B	University	Easy	Com	US	SOR	9	Organizational	31	31	Yes	0.23

a. More detailed statistics on the articles included can be found at: http://italkisee.com/data/Review/MPS/SampleDetails.xlsx. The analysis file for Comprehensive Meta-Analysis can be downloaded at: http://italkisee.com/data/Review/MPS/CMA.cma

b. B: Between-groups design; W: Within-groups design

c. PP: Paper and Pencil; Com: Computer-based

d. SOR: Structured Open Response; MC: Multiple-Choice

e. Effective sample sizes marked with # were inconsistent with the number of participants. Beveridge and Parkins (1987) reported 145 participants. The effective sample size was coded as 120, deduced from the reported event frequencies (*n* (control) = 64, *n* (experimental) = 56). Garcia-Retamero and Hoffrage's study (2013) involved 162 participants. The effective sample size was coded as 156. Garcia-Retamero et al., (2010) reported 117 participants. The effective sample size was coded as 116. Coding for both studies was based on the reported *F*-statistics and an assumption of groups with equal sample sizes to facilitate the calculation of the effect size. Lindner et al., (2016) reported 305 participants. The effective sample size was coded as the set size (*n* (control) = *n* (experimental) = 30), because this study measured item difficulty. Likewise, the effective sample size for the study by Lindner et al. (2017a) was coded as the set size (*n* (control) = *n* (experimental) = 24) rather than the reported 62 participants.

f. Hedges's *g* marked with # was combined from multiple dependent comparisons. Brase (2008) reported two dependent comparisons with one shared control group. Maries (2013) reported four comparisons from the same population. The studies by Lin et al. (2013) and Yang and Huang (2004) were based on an almost identical dataset and were therefore combined. The study by Saß et al. (2012) included four dependent comparisons with available effect sizes for representational pictures. The dependency comes from shared control or experimental groups.

were published after 2010, indicating that the multimedia effect in problem solving has attracted increasing attention in recent years. In the 51 comparisons, the total number of participants was 38,987, with a median of 63 and a range of 10 to 31,842. Excluding the study by Hoogland et al. (2018a) as an outlier (n = 31,842), the mean number of participants in the remaining studies was 143. The sets of problems used in these comparisons were primarily from science (53%) and mathematics (35%). The participants were generally primary school (43%) and university (41%) students. The number of between-groups designs (67%) was double the number of within-group designs. About half (53%) of the studies presented problems in a computer-based environment. The multiple-choice type of problem (59%) was used slightly more often than structured open responses. The average number of problems included in each study was 19, with a range of 1 to 48.

Heterogeneity analysis showed that there was a high level of heterogeneity in the outcome constructs of response accuracy (Q (33) = 783.9, I^2 = 95.79, p < .001) and response time (Q (12) = 45.553, I^2 = 73.66, p < .001) and a moderate level of heterogeneity in response certainty (Q (3) = 7.252, I^2 = 58.634, p = .064). In terms of response accuracy, heterogeneity could not be determined for decorative pictures because there was only one comparison available; comparisons between informational pictures (Q (6) = 349.396, I^2 = 98.28, p < .001) revealed the greatest heterogeneity, followed by organizational pictures (Q (7) = 41.79, I^2 = 83.25, p < .001) and representational pictures (Q (14) = 44.527, I^2 = 68.56, p < .001). The substantial amount of heterogeneity in the comparisons necessitated analysis of the possible moderators and indicated the limited reliability of the average effect sizes.

Publication Bias Analysis

The funnel plot for the response accuracy measures (k = 34) appeared to be visually symmetrical (see Fig. 5a). Furthermore, neither the regression test (t = 1.22, df = 32, p = .23) nor the rank correlation (Kendall's τ = 0.157, p = .19) showed significant asymmetry. The fail-safe N test showed that 1,302 missing publications were needed to make the significant combined effect size statistically non-significant, and it is unreasonable that so many studies go undetected. According to Rosenthal's (1991) guidelines, the fail-safe number also exceeded 5 × k + 10 (5 × k + 10 = 180 < 1,302). The overall effect size for response accuracy was therefore robust and unlikely to have been inflated by a threatening level of publication bias.

Publication bias for response accuracy was further analyzed for various types of pictures due to the considerable heterogeneity between studies. For decorative pictures (k = 1), publication bias could not be analyzed due to the lack of enough available comparisons.

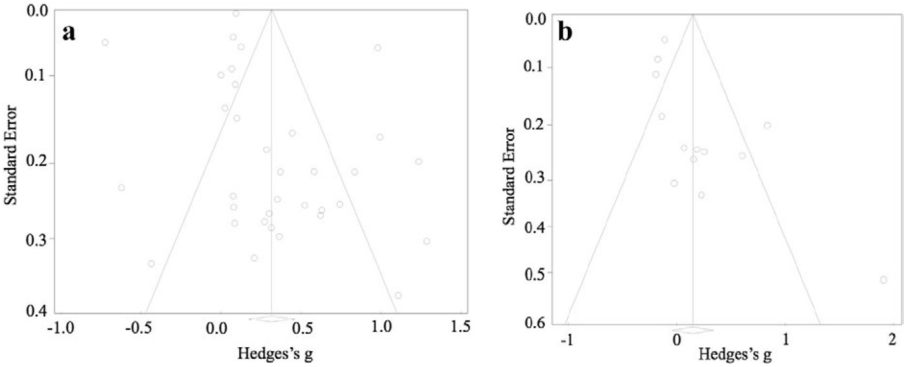


Fig. 5 Funnel plots of the outcomes. **a** Response accuracy; **b** Response time

The regression test and the rank correlation for representational ($k = 15$), organizational ($k = 8$), and informational ($k = 7$) pictures all showed no significant asymmetry. The fail-safe number for both representational ($N = 107$) and organizational ($N = 100$) pictures exceeded Rosenthal's (1991) critical values (i.e., 85 and 50, respectively), whereas for informational pictures ($N = 7$), it was below the critical value of 45. Thus, for representational and organizational pictures, there was no publication bias that could significantly threaten the validity of the average effect sizes, but publication bias could be a problem for informational pictures in measures of response accuracy.

For comparisons measuring response time ($k = 13$), the regression test ($t = 3.06$, $df = 11$, $p < .05$) and the rank correlation (Kendall's $\tau = 0.50$, $p < .05$) revealed the presence of significant publication bias. The funnel plot (see Fig. 5b) also indicated visual asymmetry. In particular, there was one extreme comparison in the lower right corner, representing the study by Goolkasian (1996). After removing this comparison ($k = 11$), the regression test ($t = 2.80$, $df = 9$, $p < .05$) and the rank correlation (Kendall's $\tau = 0.51$, $p < .05$) still indicated significant asymmetry. Therefore, the comparisons included in this study did not provide representative evidence for the multimedia effect on response time, so the aggregate effect size should be interpreted with caution.

In terms of publication bias for the effect sizes of response certainty ($k = 4$), the regression test ($t = 3.156$, $df = 2$, $p = .09$) and the rank correlation (Kendall's $\tau = 0.83$, $p = .09$) indicated no threatening publication bias. However, the fail-safe number ($N = 23$) for response certainty was below the critical value of 30 (i.e., $5 \times 4+10$), indicating that publication bias could exist. Due to the limited number of comparisons, it was difficult to evaluate the symmetry of the funnel plot. Given the inconsistent results for publication bias for response certainty, caution should also be exercised when interpreting the aggregate effect size of response certainty.

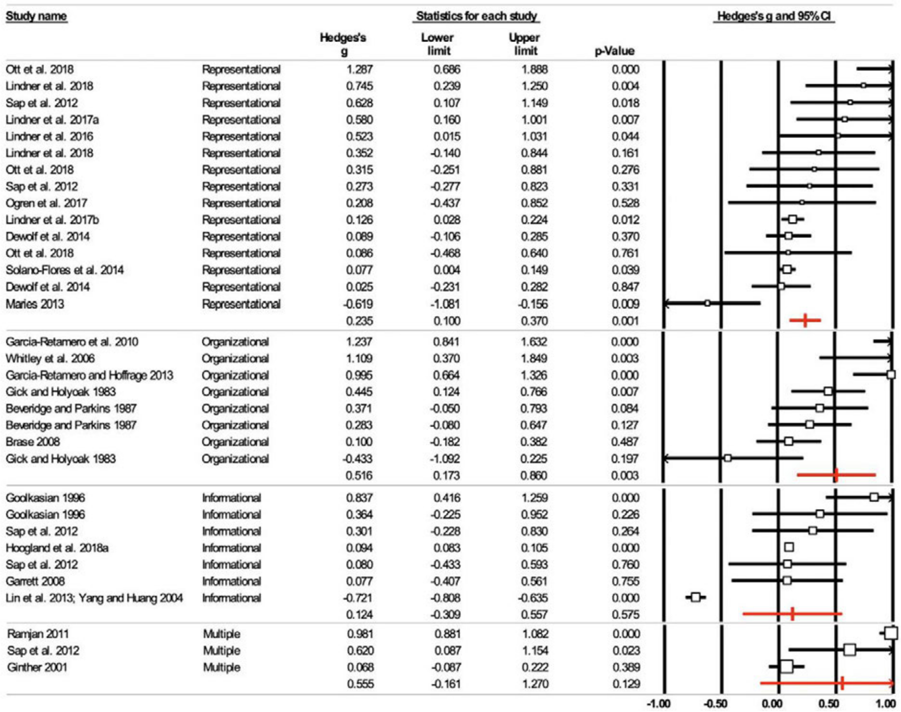
Overall Multimedia Effect in Problem Solving

Multimedia Effect on Response Accuracy

The Baujat plot for response accuracy suggested no apparent outliers. The overall estimate based on a random effects model showed that including pictures had a significant and positive average effect on response accuracy (Hedges's $g = 0.32$, $p < .001$, 95% CI [0.17, 0.46]), even after removing the two articles (Lindner et al., 2016; Lindner et al., 2017a) measuring item difficulty[1] (Hedges's $g = 0.30$, $p < .001$, 95% CI [0.16, 0.45]).

The evidence for the effect of decorative pictures was not sufficient for a reliable interpretation, as there was only one study (Berends and van Lieshout, 2009) reporting a non-significant effect on response accuracy. Without considering decorative pictures, the results showed that a picture's function did not significantly moderate the average multimedia effect on response accuracy (Q_{bet} (3) = 3.28, $p = .35$). The small aggregate effect size for representational pictures reached the significance level (Hedges's $g = 0.24$, $p < .01$, 95% CI [0.10, 0.37]). As indicated by the forest plot (see Fig. 6), Maries's (2013) study was driving the heterogeneity. Maries's study was a dissertation examining the effects of drawings in physics problem solving. Maries provided students with additional diagrams usually drawn by physics experts at the initial stage of problem solving and found that these diagrams did not scaffold students' problem solving as expected and even had a negative effect on their performance. He

[1] The constructs of *item difficulty* and *response accuracy* both refer to the outcome of response correctness, but at different levels of aggregation. Item difficulty is item focused, while response accuracy is student focused.



Study name		Statistics for each study				Hedges's g and 95% CI
		Hedges's g	Lower limit	Upper limit	p-Value	
Ott et al. 2018	Representational	1.287	0.686	1.888	0.000	
Lindner et al. 2018	Representational	0.745	0.239	1.250	0.004	
Sap et al. 2012	Representational	0.628	0.107	1.149	0.018	
Lindner et al. 2017a	Representational	0.580	0.160	1.001	0.007	
Lindner et al. 2016	Representational	0.523	0.015	1.031	0.044	
Lindner et al. 2018	Representational	0.352	-0.140	0.844	0.161	
Ott et al. 2018	Representational	0.315	-0.251	0.881	0.276	
Sap et al. 2012	Representational	0.273	-0.277	0.823	0.331	
Ogren et al. 2017	Representational	0.208	-0.437	0.852	0.528	
Lindner et al. 2017b	Representational	0.126	0.028	0.224	0.012	
Dewolf et al. 2014	Representational	0.089	-0.106	0.285	0.370	
Ott et al. 2018	Representational	0.086	-0.468	0.640	0.761	
Solano-Flores et al. 2014	Representational	0.077	0.004	0.149	0.039	
Dewolf et al. 2014	Representational	0.025	-0.231	0.282	0.847	
Maries 2013	Representational	-0.619	-1.081	-0.156	0.009	
		0.235	0.100	0.370	0.001	
Garcia-Retamero et al. 2010	Organizational	1.237	0.841	1.632	0.000	
Whitley et al. 2006	Organizational	1.109	0.370	1.849	0.003	
Garcia-Retamero and Hoffrage 2013	Organizational	0.995	0.664	1.326	0.000	
Gick and Holyoak 1983	Organizational	0.445	0.124	0.766	0.007	
Beveridge and Parkins 1987	Organizational	0.371	-0.050	0.793	0.084	
Beveridge and Parkins 1987	Organizational	0.283	-0.080	0.647	0.127	
Brase 2008	Organizational	0.100	-0.182	0.382	0.487	
Gick and Holyoak 1983	Organizational	-0.433	-1.092	0.225	0.197	
		0.516	0.173	0.860	0.003	
Goolkasian 1996	Informational	0.837	0.416	1.259	0.000	
Goolkasian 1996	Informational	0.364	-0.225	0.952	0.226	
Sap et al. 2012	Informational	0.301	-0.228	0.830	0.264	
Hoogland et al. 2018a	Informational	0.094	0.083	0.105	0.000	
Sap et al. 2012	Informational	0.080	-0.433	0.593	0.760	
Garrett 2008	Informational	0.077	-0.407	0.561	0.755	
Lin et al. 2013; Yang and Huang 2004	Informational	-0.721	-0.808	-0.635	0.000	
		0.124	-0.309	0.557	0.575	
Ramjan 2011	Multiple	0.981	0.881	1.082	0.000	
Sap et al. 2012	Multiple	0.620	0.087	1.154	0.023	
Ginther 2001	Multiple	0.068	-0.087	0.222	0.389	
		0.555	-0.161	1.270	0.129	

Fig. 6 Forest plot of the effect sizes (Hedges's *g*) of response accuracy across different function types of pictures

explained that additional diagrams led to students spending less time on conceptual planning and drawing less productive diagrams in problem solving. Maries used a quasi-experimental approach rather than a randomized controlled trial. His study also included an instruction phase in which all students learned to solve physics problems through drawing. Therefore, Maries's study may not be comparable to other studies. By removing his study, the *p* value of the significant positive average effect of representational pictures on response accuracy dropped while the average effect size slightly increased (Hedges's $g = 0.27$, $p < .001$, 95% CI [0.14, 0.39]).

Organizational pictures had a significant average effect of medium size (Hedges's $g = 0.52$, $p < .01$, 95% CI [0.17, 0.86]). Among the eight comparisons, only Gick and Holyoak (1983) reported a negative effect of organizational pictures on response accuracy. Their study focused on analogical reasoning in problem solving and examined whether an additional diagram illustrating the underlying schema of an analog story could help students solve a disparate but analogous transfer problem. They found that the participating students could interpret the diagram as representing the similar schema implicit in the two analog stories and could use the diagram as a retrieval cue to solve the transfer problem. However, they failed to understand the diagram when it was presented with only one analog story and thus performed even worse than those in the no-diagram condition. The study by Gick and Holyoak indicated that the designed organizational function may not be fulfilled if students cannot effectively understand the purpose of a picture.

Informational pictures (Hedges's $g = 0.12$, $p = .58$, 95% CI [-0.31, 0.56]) had no significant average effect on response accuracy. However, six of the seven studies showed the expected direction of the effect. As indicated by the forest plot (see Fig. 6), the combined effect size of the studies by Lin et al. (2013) and Yang and Huang (2004) was the only negative combined effect size and seemed to drive the heterogeneity of informational pictures. These two studies reported a large negative effect size of informational pictures in solving math problems for primary school students (sixth grade) in Taiwan and suggested that interpreting pictorial representations in mathematics required a specific type of mathematical ability. Multiple pictures (Hedges's $g = 0.56$, $p = .13$, 95% CI [−0.16, 1.27]) also had no significant aggregate effect, although the three included comparisons reported positive effects on response accuracy.

Problem difficulty (Q_{bet} (1) = 0.14, p = .71) was not a significant moderator of the multimedia effect on response accuracy. Adding pictures to difficult (Hedges's $g = 0.28$, $p <$ .01, 95% CI [0.11, 0.46]) or easy problems (Hedges's $g = 0.34$, $p < .01$, 95% CI [0.08, 0.60]) both significantly improved an individual's response accuracy with small-to-medium average effect sizes. Other possible moderators were also tested (see Table 3). They were all found to be non-significant, including institution location (Q_{bet} (2) = 2.86, p = .24), effective sample size (Q_M (1) = 0.31, p = .58), set size (Q_M (1) = 0.56, p = .46), problem domain (Q_{bet} (3) = 7.09, p = .07), answer format (Q_{bet} (1) = 0.02, p = .90), test environment (Q_{bet} (1) = 0.69, p = .41),

Table 3 Moderator analysis of the multimedia effect on response accuracya

Possible moderators	k	Heterogeneity		Average effect size		
		Q	I^2	Hedges's g	95% CI	p
Institution location						
(1) Europe***	22	104.4	79.88%	0.36	[0.24, 0.49]	< .001
(2) United States	10	36.49	75.33%	0.17	[−0.02, 0.36]	0.07
(3) Other	2	635.8	99.84%	0.13	[−1.54, 1.80]	0.88
Effective sample size	29	0.31	—	—	—	0.58
Set size	33	0.56	—	—	—	0.46
Problem difficulty						
(1) Easy**	22	730.5	97.13%	0.34	[0.08, 0.60]	< .01
(2) Difficult**	12	53.42	79.41%	0.28	[0.11, 0.46]	< .01
Domain						
(1) Science**	17	47.82	66.54%	0.28	[0.15, 0.42]	< .001
(2) Math	12	672.5	98.36%	0.27	[−0.04, 0.59]	0.09
(3) Medicine*	3	27.21	92.65%	0.77	[0.06, 1.48]	< .05
(4) Language	2	0.001	0.00%	0.07	[-0.08, 0.22]	0.36
Answer format						
(1) Multiple-choice**	19	334.8	94.62%	0.33	[0.09, 0.58]	< .01
(2) Structured open response**	15	385.1	96.37%	0.31	[0.08, 0.54]	< .01
Test environment						
(1) Computer-based***	14	66.33	80.40%	0.38	[0.24, 0.53]	< .001
(2) Paper and pencil	20	716.9	97.35%	0.25	[−0.05, 0.53]	0.10
Education level						
(1) Primary	14	253.4	94.87%	0.18	[−0.09, 0.46]	0.20
(2) University**	13	96.39	87.55%	0.47	[0.16, 0.78]	< .01
(3) Other***	7	67.84	91.16%	0.32	[0.15, 0.48]	< .001
Experimental design						
(1) Within-groups design	11	665	98.50%	0.23	[−0.004, 0.47]	0.054
(2) Between-groups design***	23	93.14	76.38%	0.36	[0.18, 0.55]	< .001
Control conditions						
(1) Without control variables	10	681.8	98.68%	0.21	[−0.18, 0.60]	0.29
(2) With single or multiple control variables***	24	101	77.24%	0.35	[0.21, 0.48]	< .001

[a] k = number of pairwise comparisons. Q = weighted squared deviations of the aggregate effect size within each subgroup. I^2 = percentage of unexplained variance in aggregate effect size. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

participants' education level (Q_{bet} (2) = 1.81, p = .41), experimental design (Q_{bet} (1) = 0.72, p = .40), and control conditions (Q_{bet} (1) = 0.40, p = .53).

Multimedia Effect on Response Time

Goolkasian's (1996) study was found to be a potential outlier for response time. There were two qualified comparisons in this study, both of which were influential, with large residuals (externally standardized residuals were 2.12 and 3.09, respectively) and a strong influence on the results. Removing Goolkasian's study significantly reduced the heterogeneity (from Q (12) = 45.55 to Q (10) = 13.03). Goolkasian (1996) provided students with pictorial or textual background information, then asked them to respond to the test sentences (true/false) as quickly as possible. Compared with other studies, he reported a much greater reduction in response time when students were provided with pictures. After removing Goolkasian's study, the meta-analysis did not find a significant and positive average multimedia effect on response time (Hedges's g = −0.06, p = .31, 95% CI [−0.16, 0.05]) though 6 of the 11 studies showed a positive multimedia effect on response time (see Fig. 7). The effect was still non-significant when Goolkasian's study was included (Hedges's g = 0.15, p = .13, 95% CI [−0.04, 0.34]). Therefore, the multimedia effect on response time remained uncertain based on these limited and inconsistent findings.

Due to the limited number of comparisons, the moderating effect of a picture's function type was not tested. For decorative pictures, there was only one comparison available (see Fig. 7). Berends and van Lieshout (2009) reported that decorative pictures significantly increased the response time. Three of the five comparisons also indicated that representational pictures increased response time. In addition, only one negative effect (Lindner et al., 2017b) among the five comparisons reached the significance level. The only comparison for organizational pictures reported a non-significant and positive multimedia effect on response time (Whitley et al., 2006). For informational pictures, four (Goolkasian, 1996; Saß et al., 2012) of the five comparisons suggested their significant effects in reducing response time. For multiple pictures, there was only one comparison available (Saß et al., 2012), which indicated a positive but non-significant multimedia effect on response time.

Due to the limited number of comparisons, moderator analysis was also not conducted for problem difficulty. After removing Goolkasian's (1996) study, response time was found to be significantly moderated by test environment (Q_{bet} (1) = 4.39, p < .05) and experimental design

Study name	Function Type	Statistics for each study				Relative weight
		Hedges's g	Lower limit	Upper limit	p-Value	
Berends and van Lieshout 2009	Decorative	-0.173	-0.345	-0.000	0.049	21.14
Lindner et al. 2017a	Representational	-0.134	-0.523	0.255	0.499	6.60
Lindner et al. 2017b	Representational	-0.108	-0.206	-0.010	0.031	33.20
Ogren et al. 2017	Representational	-0.021	-0.663	0.622	0.950	2.66
Sap et al. 2012	Representational	0.158	-0.394	0.710	0.574	3.53
Sap et al. 2012	Representational	0.071	-0.437	0.579	0.785	4.11
Whitley et al. 2006	Organizational	0.230	-0.459	0.918	0.513	2.33
Sap et al. 2012	Informational	0.187	-0.327	0.701	0.475	4.02
Schwert 2007	Informational	-0.189	-0.419	0.042	0.109	14.81
Sap et al. 2012	Informational	0.606	0.068	1.144	0.027	3.70
Sap et al. 2012	Multiple	0.257	-0.266	0.780	0.336	3.89
		-0.056	-0.164	0.052	0.310	

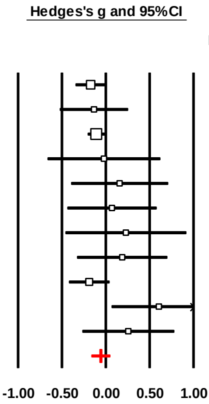


Fig. 7 Forest plot of the effect sizes (Hedges's g) of response time

(Q_{bet} (1) = 9.42, p < .01). Additional pictures in problems using a paper-and-pencil format significantly increased response time (Hedges's g = −0.18, p < .05, 95% CI [−0.32, −0.04]), whereas a computer-based environment had a non-significant average effect (Hedges's g = 0.05, p = .57, 95% CI [−0.11, 0.21]). For studies adopting a between-subjects design, a significant average multimedia effect on response time was detected (Hedges's g = 0.22, p < .05, 95% CI [0.01, 0.43]), whereas within-subject designs showed a significant reverse effect (Hedges's g = −0.13, p < .01, 95% CI [−0.21, −0.05]).

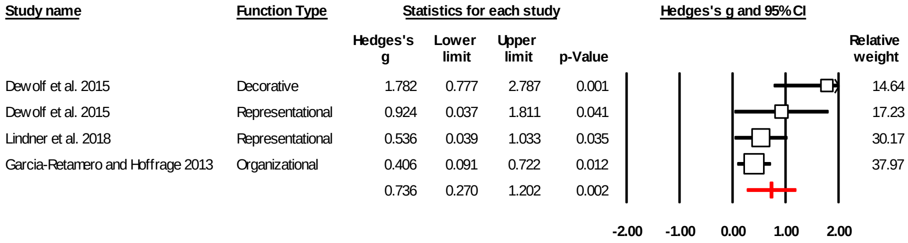
Across the included studies, there was no significant moderating effect of institution location (Q_{bet} (1) = 0.23, p = .64), effective sample size (Q *(1)* = 1.14, p = .28), set size (Q *(1)* = 0.14, p = .70), problem domain (Q_{bet} (1) = 2.12, p = .15), answer format (Q_{bet} (1) = 3.53, p = .06), participants' education level (Q_{bet} (1) = 0.84, p = .36) or control conditions (Q_{bet} (1) = 3.03, p = .08). The presence of pictures significantly increased the response time when studies had no control variables (Hedges's g = −0.12, p < .01, 95% CI [−0.21, −0.03]) or adopted items requiring structured open responses (Hedges's g = −0.16, p < .05, 95% CI [−0.30, −0.03]).

Multimedia Effect on Metacognition

There were four comparisons for response certainty. It should be noted that Garcia-Retamero and Hoffrage (2013) measured perceived task difficulty, which was assumed to be consistent with response certainty. The analysis revealed a positive and significant aggregate effect size for response certainty (Hedges's g = 0.74, p < .01, 95% CI [0.27, 1.20]) (see Fig. 8). However, the limited number of comparisons may affect the reliability of this result. In addition, no moderation analysis could be conducted due to insufficient comparisons.

Discussion

This meta-analysis showed a significant small-to-medium average multimedia effect on response accuracy (Hedges's g = 0.32) and a significant medium-to-large average multimedia effect on students' response certainty (Hedges's g = 0.74), but no significant multimedia effect on response time on average. Overall, we found evidence to support Hypotheses 1 and 2. However, the limited data did not allow for a clear evaluation of these two hypotheses and the multimedia effect on response time.



Study name	Function Type	Statistics for each study				Hedges's g and 95% CI
		Hedges's g	Lower limit	Upper limit	p-Value	Relative weight
Dewolf et al. 2015	Decorative	1.782	0.777	2.787	0.001	14.64
Dewolf et al. 2015	Representational	0.924	0.037	1.811	0.041	17.23
Lindner et al. 2018	Representational	0.536	0.039	1.033	0.035	30.17
Garcia-Retamero and Hoffrage 2013	Organizational	0.406	0.091	0.722	0.012	37.97
		0.736	0.270	1.202	0.002	

Fig. 8 Forest plot of the effect sizes (Hedges's g) of response certainty

Multimedia Effect on Response Accuracy

Overall, the significant multimedia effect in problem solving in the included studies is consistent with the multimedia effect in learning (Butcher, 2014; Mayer, 2001, 2009, 2014). Early meta-analyses of the multimedia effect in learning have shown that decorative pictures are not beneficial for text learning, whereas other types of pictures can enhance text-based learning with medium-to-large effect sizes (Levin et al., 1987). Recent studies on the multimedia effect in learning have also reported medium-to-large effect sizes (Butcher, 2014). However, variations in the problem-solving context complicate and diversify the multimedia effect. The 34 pairwise comparisons on response accuracy included in this meta-analysis showed high heterogeneity. However, it did not identify important moderators.

Moderating Effect of Function Type

For decorative pictures, there was only one comparison available indicating a non-significant effect on response accuracy (Berends and van Lieshout, 2009). This somewhat echoes the double-edged effects of seductive details or decorative pictures on multimedia learning outcomes. On the one hand, seductive details or decorative pictures are task-irrelevant details and may thus hinder learning when they induce overload (Lehman et al., 2007), unless students can intentionally ignore them, which depends on their task experience (Rop, 2017; Rop et al., 2017) or their perceived irrelevance of these seductive details (Eitel et al., 2019). On the other hand, seductive details or decorative pictures capture very little attention (Lenzner et al., 2013), but help elicit positive affective experiences and lower perceived cognitive load, which may promote germane processing when cognitive resources are available (Park et al., 2011; Schneider et al., 2016).

By not considering decorative pictures, the function of the pictures did not significantly moderate the multimedia effect on response accuracy. Adding representational (Hedges's g = 0.24) and organizational (Hedges's g = 0.52) pictures significantly improved the response accuracy based on the included studies, but this effect was not found for informational or multiple pictures. The overall results of this study cannot fully confirm the previous findings that task-relevant pictures are beneficial for response accuracy while decorative pictures have little or no effect on it (Agathangelou et al., 2008; Carney and Levin, 2002; Elia et al., 2007; Lindner et al., 2018). Pictures with different functions had varying degrees of the multimedia effect on response accuracy, which is difficult to interpret. This study presents some alternative explanations for the effects of organizational and informational pictures.

Regarding the strong positive effect of organizational pictures, the DeFT (Design, Functions, Tasks) framework proposed by Ainsworth (2006) can offer some insights. This framework suggests that various combinations of external representations mainly fulfill three key functions: generating complementary processes or providing complementary information in mental model construction; constraining interpretations of any representation; and constructing a deeper understanding through an integration process. According to the DeFT framework, decorative pictures have no prominent functions; representational pictures tend to complement and constrain interpretations of corresponding texts; informational pictures and texts tend to complement each other by providing unique information; and organizational pictures are most likely to induce all types of functions and thus best support mental model construction.

The significant impact of organizational pictures can also be explained by the difference between pictures and texts. Pictures are assumed to be more computationally efficient than texts in conveying spatial information. Thus, organizational pictures make full use of the picture advantage to provide spatial information and scaffold mental model construction. This is consistent with a recent finding in learning that overlapping spatial information is beneficial, but overlapping non-spatial information has no effect on learning performance (Schüler et al., 2019).

Informational pictures are quite common in problem-solving contexts (Dewolf, 2014). Thus, the non-significant aggregate effect of informational pictures is worth discussing, as well as why they sometimes inhibit problem solving (Lin et al., 2013; Maries, 2013). Informational pictures replace part of the problem text with an equivalent pictorial format. Maries (2013) attributed the detrimental effect of informational pictures to students spending less time on conceptual analysis of the problem.

The text-driven assumption may provide an alternative explanation. There has been plenty of evidence showing that text is a dominant cue in mental model construction and guide picture processing (e.g., Canham and Hegarty, 2010; Folker et al., 2005; Hao, 2010; Hegarty and Just, 1993; Schüler et al., 2019). Therefore, informational pictures may not be fully processed due to the lack of textual guidance, as they have little overlap with the problem text.

Another explanation concerns an individual's representational competence. Informational pictures contain information that is indispensable for problem solving, and thus usually complement the text. The ability to interpret informational pictures is therefore necessary to solve such problems. Students with low levels of representational competence (i.e., the ability to use representations to understand and communicate ideas and solve problems [Daniel et al., 2018; Huinker, 2015]) may find it difficult to process pictorial information. Students who have difficulty integrating pictorial and descriptive representations may find it more difficult to solve text–picture problems.

The way in which informational pictures are combined with the problem text also affects student performance. The text–picture format has been shown to be more important for problems with informational pictures than for those with other types of pictures (Saß et al., 2017). An integrated format is needed for complementary information representations to support individual problem solving. An individual's test performance will be hampered when relevant content is presented in a split manner (Jarodzka et al., 2015).

Multimedia Effect on Response Time

The results of this meta-analysis did not show a significant average multimedia effect on response time. The benefits of the picture superiority effect (Nickerson, 1965; Shepard, 1967) and the dual processing assumption (Schneider and Chein, 2003; Shiffrin and Schneider, 1977) do not appear to be strong enough to generate a consistent change in response time when equivalent pictures are added to illustrate or replace part of the problem text. However, this finding should be interpreted with caution due to the presence of publication bias, the very limited number of primary studies, and a high level of heterogeneity in the measures of response time.

Pictures with various functions affect response time in a very complex way. Eye-tracking research has shown that students pay different attention to different types of pictures (Dewolf et al., 2015; Lindner et al., 2017a). However, this study could not effectively separate the average multimedia effect on response time induced by different types of pictures given the small number of studies. Therefore, future primary studies should measure response time and adopt techniques such as eye tracking or thinking aloud to further examine picture processing in problem solving.

Whether pictures can function as recall cues also influences response time. For example, Saß et al., (2012) found that pictures in the stems and answer options of multiple-choice items can significantly reduce students' response time. They suggested that adding pictures to answer options can facilitate solution processing through analog structure mapping processes (Schnotz and Bannert, 2003). This is consistent with the results of previous studies on analogical problem solving in which pictures in analog items serve as recall cues for information retrieval (Beveridge and Parkins, 1987; Gentner et al., 2003). This visual analog effect in the stems and answer options also applies to study and test contexts in which a matching picture in both study materials and test items can improve test performance (Lindner et al., 2018). This is also referred to as *transfer appropriate processing* in recognition tasks, which suggests that memory is improved when there is a match rather than a mismatch between item processing for a test and for study (e.g., Morris et al., 1977).

Readers should also be cautious with the interpretation of response time itself due to its heterogeneous effects on an individual's problem-solving performance (Goldhammer et al., 2014; Goldhammer et al., 2015). There is no clear answer as to whether a lower response time is good or bad. On the one hand, students' response time may indicate their speed of information processing (Scherer et al., 2015). On the other hand, a reduction in response time may indicate a reduced level of engagement such as rapid-guessing behavior. These two-sided effects are also moderated by task difficulty and individual skill levels (Goldhammer et al., 2014; Goldhammer et al., 2015). For difficult problems or low-skilled individuals, a longer response time is more likely to indicate better performance. Consequently, simply using response time may be inadequate for measuring multimedia problem-solving performance if the response time itself is not an evaluation criterion.

Multimedia Effect on Metacognition

The multimedia effect on an individual's metacognition in problem solving was supported in this study, but this result may be affected by publication bias. The students perceived greater response certainty when solving text–picture problems than when solving text-only problems (Hedges's $g = 0.74$). The higher metacognitive judgment induced by pictures has also been observed in learning contexts, in which seductive details reduce students' perceived cognitive load and task difficulty (Lenzner et al., 2013; Magner et al., 2014; Park et al., 2011; Schneider et al., 2016). Higher metacognitive judgments under text–picture problem conditions may also benefit students' positive affective experiences (e.g., motivation, self-efficacy, pleasure, arousal) in multimedia problem solving.

However, increased confidence in problem solving induced by pictures may have detrimental effects when metacognitive judgments are biased. Phenomena such as overconfidence and underestimated task difficulty are commonly referred to as metacognitive illusions or bias due to the use of multimedia materials in both problem solving (Ögren et al., 2017) and learning contexts (Bjork et al., 2013; Eitel, 2016; Serra and Dunlosky, 2010; Wiley et al., 2014). For example, students tend to confirm statements accompanied by pictures and perceive conflicting text information as plausible (Isberner et al., 2013; McCabe and Castel, 2008; Ögren et al., 2017). Inflated metacognitive judgments may also lead to suboptimal restudy behaviors and less mental effort in text comprehension (Lindner et al., 2018; Ögren et al., 2017). Consequently, the question of whether an individual's increased confidence due to pictures in problems is biased should be carefully considered.

Limitations and Directions for Future Research

This meta-analysis has several limitations and its results should be interpreted with particular caution. First, the quality of the studies included in the sample cannot be fully guaranteed. Due to the current limited number of quantitative studies on this topic, we adopted a broad literature search strategy without strict exclusion criteria in terms of experimental design or publication source. Moreover, publication bias could exist in the multimedia effect on response accuracy (for informational pictures), response time, and metacognition. Future review research should better address these questions by including more stringent quality evaluation criteria when selecting the literature.

Second, the limited number of studies and high levels of heterogeneity may also have influenced the findings. Due to limited studies on this topic, only 26 articles were qualified based on our criteria. These studies came from different fields, their participants ranged from primary students to adults and they had different sample sizes. The heterogeneity of response accuracy and response time was far too high. A high level of heterogeneity may indicate a lack of precision in aggregate effect sizes. Therefore, a random effects model was adopted. Moderator analysis was also conducted to decompose this heterogeneity. However, we did not find significant moderators for response accuracy, leaving the heterogeneity still unexplained.

This study estimated problem difficulty based on students' response accuracy on text problems, because most studies did not report the difficulty level of their items. Therefore, it should be noted that the problem difficulty code in this study may not capture the actual difficulty level. In addition, due to limited data, this study did not examine other important moderators identified in the literature, such as the text–picture integration format, an individual's prior knowledge, and problem complexity. Future review research should test and interpret these other possible moderating effects when more qualified studies are published.

Finally, this study only categorized pictures according to their functional relationship with the problem texts. However, the main categorization of picture types is not always perfectly clear. In addition, it is difficult to determine whether a certain type of picture was interpreted by the students as it was designed to be. At the same time, pictures identified as having the same function may vary considerably. For example, representational pictures may include cartoons or realistic photographs. The abstractness of a picture may also moderate the multimedia effect in problem solving by invoking different contexts or levels of detail. This study did not categorize pictures based on this dimension. This limitation should encourage further research on the moderating effect of picture type.

Conclusion and Implications

This meta-analytic review supports the multimedia effect in problem solving to some extent and highlights the impact of pictures on problem-solving performance in common exercises and formal tests. The use of multiple media in assessments or problem-based learning raises a number of issues that have not yet been fully explored. On the one hand, multimedia can improve item access and serve as an accommodation to minimize the influence of language proficiency or reading comprehension skills (Solano-Flores et al., 2014; Winter et al., 2006). On the other hand, an individual's ability to interpret multiple representations is likely to cause access issues if representational competence is not included in the targeted measurement construct. This review also casts doubt on the direct application of multimedia learning principles to guide item design. Kirschner et al. (2017) proposed a cognitive theory of multimedia assessment that reversed most of the original multimedia learning principles to better differentiate novices and experts and ensure ecological validity. However, these claims remain empirically untested. Further prospective studies are needed to investigate item design principles in the contexts of problem-based learning, typical exercises, and formal testing.

Acknowledgements We greatly appreciate the support and advice from the editor and anonymous reviewers. We would also like to thank Ms. Xiao, Nan from the University of Hong Kong for her precious comments on an earlier version of the manuscript. This research was supported by Hong Kong RGC grant no. 27606915 and Hong Kong PICO grant no. 2017.A8.073.18C.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Agathangelou, S., Gagatsis, A., & Papakosta, V. (2008). The role of verbal description, representational and decorative picture in mathematical problem solving. In A. Gagatsis (Ed.), *Research in mathematics education: Conference of five cities: Nicosia, Rhodes, Bologna, Palermo, Locarno* (pp. 39–56). Cyprus: University of Cyprus.

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

Baddeley, A. D. (1992). Working memory. *Science*, 255(5044), 556–559.

Baddeley, A. D., Allen, R. J., & Hitch, G. J. (2011). Binding in visual working memory: the role of the episodic buffer. *Neuropsychologia*, 49(6), 1393–1400. https://doi.org/10.1016/j.neuropsychologia.2010.12.042.

Baujat, B., Mahé, C., Pignon, J. P., & Hill, C. (2002). A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Statistics in Medicine*, 21(18), 2641–2652.

Berends, I. E., & van Lieshout, E. C. D. M. (2009). The effect of illustrations in arithmetic problem-solving: effects of increased cognitive load. *Learning and Instruction*, 19(4), 345–353. https://doi.org/10.1016/j.learninstruc.2008.06.012.

Beveridge, M., & Parkins, E. (1987). Visual representation in analogical problem solving. *Memory and Cognition*, 15(3), 230–237.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology*, 64(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823.

Bodner, M. G., & Domin, D. S. (2000). Mental models: the role of representations in problem solving in chemistry. *University Chemistry Education*, 4(1), 24–30.

Boonen, A. J. H., van Wesel, F., Jolles, J., & van der Schoot, M. (2014). The role of visual representation type, spatial ability, and reading comprehension in word problem solving: an item-level analysis in elementary school children. *International Journal of Educational Research*, 68, 15–26.

Brase, G. L. (2008). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, 23(3), 369–381.

Butcher, K. R. (2014). The multimedia principle. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 174–205). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.010.

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction*, 20(2), 155–166. https://doi.org/10.1016/j.learninstruc.2009.02.014.

Carney, R., & Levin, J. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*, 14(1), 5–26. https://doi.org/10.1023/A:1013176309260.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4), 293–332.

Chuah, J, Zhang, J, & Johnson, T. R. (2000). The representational effect in complex systems: a distributed representation approach. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 633–638). Hillsdale: Erlbaum.

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research*, 48(2), 139–154. https://doi.org/10.1080/00131880600732249.

Daniel, K. L., Bucklin, C. J., Leone, E. A., & Idema, J. (2018). Towards a definition of representational competence. In K. L. Daniel (Ed.), *Towards a framework for representational competence in science education* (pp. 3–11). Cham: Springer.

Dewolf, T. (2014). *Get the picture? Are representational illustrations effective in helping pupils to solve mathematical word problems realistically?* University of Leuven: Unpublished doctoral dissertation.

Dewolf, T., Van Dooren, W., Ev Cimen, E., & Verschaffel, L. (2014). The impact of illustrations and warnings on solving mathematical word problems realistically. *Journal of Experimental Education*, 82(1), 103–120. https://doi.org/10.1080/00220973.2012.745468.

Dewolf, T., Van Dooren, W., Hermens, F., & Verschaffel, L. (2015). Do students attend to representational illustrations of non-standard mathematical word problems, and, if so, how helpful are they? *Instructional Science*, 43(1), 147–171. https://doi.org/10.1007/s11251-014-9332-7.

Dewolf, T., Van Dooren, W., & Verschaffel, L. (2017). Can visual aids in representational illustrations help pupils to solve mathematical word problems more realistically? *European Journal of Psychology of Education*, 32(3), 335–351. https://doi.org/10.1007/s10212-016-0308-7.

Dindar, M., Yurdakul, I. K., & Dönmez, F. I. (2013). Multimedia in test items: animated questions vs. static graphics questions. *Procedia - Social and Behavioral Sciences*, 106, 1876–1882. https://doi.org/10.1016/j.sbspro.2013.12.213.

Eitel, A. (2016). How repeated studying and testing affects multimedia learning: evidence for adaptation to task demands. *Learning and Instruction*, 41, 70–84. https://doi.org/10.1016/j.learninstruc.2015.10.003.

Eitel, A., Bender, L., & Renkl, A. (2019). Are seductive details seductive only when you think they are relevant? An experimental test of the moderating role of perceived relevance. *Applied Cognitive Psychology*, 33(1), 20–30. https://doi.org/10.1002/acp.3479.

Elia, I., & Philippou, G. (2004). The functions of pictures in problem solving. In M. J. Hoines & A. B. Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 327–334). Bergen: PME.

Elia, I., Gagatsis, A., & Demetriou, A. (2007). The effects of different modes of representation on the solution of one-step additive problems. *Learning and Instruction*, 17(6), 658–672. https://doi.org/10.1016/j.learninstruc.2007.09.011.

Folker, S., Ritter, H., & Sichelschmidt, L. (2005). Processing and integrating multimodal material—the influence of color-coding. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 690–695). Mahwah: Erlbaum.

Fuchs, L. S., Fuchs, D., Compton, D. L., Hamlett, C. L., & Wang, A. Y. (2015). Is word-problem solving a form of text comprehension? *Scientific Studies of Reading*, 19(3), 204–223. https://doi.org/10.1080/10888438.2015.1005745.

Gagatsis, A., & Elia, E. (2004). The effects of different modes of representation on mathematical problem solving. In M. J. Hoines & A. B. Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group of the Psychology of Mathematics Education* (Vol. 2, pp. 447–454). Bergen: PME.

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science and Medicine*, 83, 27–33.

Garcia-Retamero, R., Galesic, M., & Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? *Medical Decision Making*, 30(6), 672–684. https://doi.org/10.1177/0272989X10369000.

Garrett, A. J. (2008). *The role of picture perception in children's performance on a picture vocabulary test* (Doctoral dissertation). Retrieved from ProQuest Dissertations Publishing. (Accession No. 3324639).

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: a general role for analogical encoding. *Journal of Educational Psychology*, 95(2), 393–408.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. New York: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. New York: Houghton Mifflin.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15(1), 1–38. https://doi.org/10.1016/0010-0285(83)90002-6.

Ginther, A. (2001). *Effects of the presence and absence of visuals on performance on TOEFL CBT listening-comprehension stimuli. (TOEFL research report no. 66)*. Princeton: Educational Testing Service.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106(3), 608–626. https://doi.org/10.1037/a0034716.

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: the relation between item response and item response time in raven's matrices. *Journal of Intelligence*, 3(1), 21–40. https://doi.org/10.3390/jintelligence3010021.

Goolkasian, P. (1996). Picture-word differences in a sentence verification task. *Memory and Cognition*, 24(5), 584–594. https://doi.org/10.3758/bf03201085.

Greco, T., Zangrillo, A., Biondi-Zoccai, G., & Landoni, G. (2013). Meta-analysis: pitfalls and hints. *Heart, Lung and Vessels*, 5(4), 219–225. https://doi.org/10.4028/www.scientific.net/AMR.60-61.110.

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of complex problem solving: concept, implementation, and application. *Educational Technology Research and Development*, 61(3), 407–421. https://doi.org/10.1007/s11423-013-9301-x.

Hao, Y. (2010). Does multimedia help students answer test items? *Computers in Human Behavior*, 26(5), 1149–1157. https://doi.org/10.1016/j.chb.2010.03.021.

Hardy-Vallée, B., & Payette, N. (Eds.). (2009). *Beyond the brain: embodied, situated and distributed cognition*. Newcastle: Cambridge Scholars Publishing.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: a theory of cognitive interest in science learning. *Journal of Educational Psychology*, 90(3), 414–434. https://doi.org/10.1037/0022-0663.90.3.414.

Hartmann, S., & Leutner, D. (2013). Die Rolle von Leseverständnis und Lesegeschwindigkeit beim Zustandekommen der Leistungen in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenz (The role of reading comprehension and reading speed in text-based assessments of scientific inquiry skills). (Doctoral dissertation, University of Duisburg-Essen). Retrieved from http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-33260/hartmann_diss.pdf.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language*, 32(6), 717–742. https://doi.org/10.1006/jmla.1993.1036.

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

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(2), 174–196.

Hoogland, K., de Koning, J., Bakker, A., Pepin, B. E. U., & Gravemeijer, K. (2018a). Changing representation in contextual mathematical problems from descriptive to depictive: the effect on students' performance. *Studies in Educational Evaluation*, 58(June), 122–131. https://doi.org/10.1016/j.stueduc.2018.06.004.

Hoogland, K., Pepin, B., de Koning, J., Bakker, A., & Gravemeijer, K. (2018b). Word problems versus image-rich problems: an analysis of effects of task characteristics on students' performance on contextual mathematics problems. *Research in Mathematics Education*, 20(1), 37–52. https://doi.org/10.1080/14794802.2017.1413414.

Huinker, D. (2015). Representational competence: a renewed focus for classroom practice in mathematics. *Wisconsin Teacher of Mathematics*, 67(2), 4–8.

Isberner, M. B., Richter, T., Maier, J., Knuth-Herzig, K., Horz, H., & Schnotz, W. (2013). Comprehending conflicting science-related texts: graphs as plausibility cues. *Instructional Science*, 41(5), 849–872. https://doi.org/10.1007/s11251-012-9261-2.

Jarodzka, H., Janssen, N., Kirschner, P. A., & Erkens, G. (2015). Avoiding split attention in computer-based testing: is neglecting additional information facilitative? *British Journal of Educational Technology*, 46(4), 803–817. https://doi.org/10.1111/bjet.12174.

Johnson-Laird, P. N. (2005). Mental models and thought. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 185–208). New York: Cambridge University Press.

Johnstone, A. H. (1982). Macro and microchemistry. *School Science Review*, 64, 377–379.

Kirschner, P., Park, B., Malone, S., & Jarodzka, H. (2017). Towards a cognitive theory of multimedia assessment (CTMMA). In J. M. Spector, B. B. Lockee, & M. Childress (Eds.), *Learning, design, and technology: an International compendium of theory, research, practice, and policy* (pp. 1–23). Cham: Springer.

Lehman, S., Schraw, G., McCrudden, M. T., & Hartley, K. (2007). Processing and recall of seductive details in scientific text. *Contemporary Educational Psychology*, 32(4), 569–587.

Lenzner, A., Schnotz, W., & Müller, A. (2013). The role of decorative pictures in learning. *Instructional Science*, 41(5), 811–831. https://doi.org/10.1007/s11251-012-9256-z.

Lesh, R., Post, T., & Behr, M. (1987). Representations and translations among representations in mathematics learning and problem solving. In C. Janvier (Ed.), *Problems of representations in the teaching and learning of mathematics* (pp. 33–40). Hillsdale: Lawrence Erlbaum.

Levin, J. R. (1981). On the functions of pictures in prose. In F. J. Pirozzolo & M. C. Wittrock (Eds.), *Neuropsychological and cognitive processes in reading* (pp. 203–228). San Diego: Academic Press.

Levin, J. R., Anglin, G. J., & Carney, R. N. (1987). On empirically validating functions of pictures in prose. In D. M. Willows & H. A. Houghton (Eds.), *The psychology of illustration: Vol. 1. Basic research* (pp. 51–85). New York: Springer-Verlag.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: the science of reviewing research*. Cambridge: Harvard University Press.

Lin, Y.-H., Wilson, M., & Cheng, C.-L. (2013). An investigation of the nature of the influences of item stem and option representation on student responses to a mathematics test. *European Journal of Psychology of Education*, 28(4), 1141–1161. https://doi.org/10.1007/s10212-012-0159-9.

Lindner, M. A., Ihme, J. M., Saß, S., & Köller, O. (2016). How representational pictures enhance students' performance and test-taking pleasure in low-stakes assessment. *European Journal of Psychological Assessment*, 34(6), 376–385. https://doi.org/10.1027/1015-5759/a000351.

Lindner, M. A., Eitel, A., Strobel, B., & Köller, O. (2017a). Identifying processes underlying the multimedia effect in testing: an eye-movement analysis. *Learning and Instruction*, 47, 91–102. https://doi.org/10.1016/j.learninstruc.2016.10.007.

Lindner, M. A., Lüdtke, O., Grund, S., & Köller, O. (2017b). The merits of representational pictures in educational assessment: evidence for cognitive and motivational effects in a time-on-task analysis. *Contemporary Educational Psychology*, 51, 482–492. https://doi.org/10.1016/j.cedpsych.2017.09.009.

Lindner, M. A., Eitel, A., Barenthien, J., & Köller, O. (2018). An integrative study on learning and testing with multimedia: effects on students' performance and metacognition. *Learning and Instruction*, 71, 101100. https://doi.org/10.1016/j.learninstruc.2018.01.002.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.

Magner, U. I. E., Schwonke, R., Aleven, V., Popescu, O., & Renkl, A. (2014). Triggering situational interest by decorative illustrations both fosters and hinders learning in computer-based learning environments. *Learning and Instruction*, 29, 141–152. https://doi.org/10.1016/j.learninstruc.2012.07.002.

Maries, A. (2013). *Role of multiple representations in physics problem solving* (Doctoral dissertation). Retrieved May 21, 2019 from http://d-scholarship.pitt.edu/20000/1/Alex_Thesis_ETD6.pdf

Martin, M. O., Mullis, I. V. S., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in science*. International Association for the Evaluation of Educational Achievement. Retrieved May 21, 2019 from http://timssandpirls.bc.edu/timss2015/international-results/wp-content/uploads/filebase/full%20 pdfs/T15-International-Results-in-Science-Grade-8.pdf

Mayer, R. E. (1987). *Educational psychology: a cognitive approach*. Boston: Little, Brown.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York: Cambridge University Press.

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 31–48). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.005.

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: the effect of brain images on judgments of scientific reasoning. *Cognition*, 107(1), 343–352. https://doi.org/10.1016/j.cognition.2007.07.017.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Reprint-preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Physical Therapy*, 89(9), 873–880.

Moreno, R. (2006). Does the modality principle hold for different media? A test of the method-affects-learning hypothesis. *Journal of Computer Assisted Learning*, 22(3), 149–158. https://doi.org/10.1111/j.1365-2729.2006.00170.x.

Moreno, R. (2009). Learning from animated classroom exemplars: the case for guiding student teachers' observations with metacognitive prompts. *Educational Research and Evaluation*, 15(5), 487–501. https://doi.org/10.1080/13803610903444592.

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, 19(3), 309–326. https://doi.org/10.1007/s10648-007-9047-2.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16(5), 519–533.

Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction*, 9(4), 329–389.

Newell, A. (1990). *Unified theories of cognition*. Cambridge: Harvard University Press.

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: a demonstration of capacity. *Canadian Journal of Psychology*, 19(2), 155–160. https://doi.org/10.1037/h0082899.

OECD. (2007). *PISA 2006: Science competencies for tomorrow's world: Volume 1: analysis. Paris: PISA, OECD Publishing.* https://doi.org/10.1787/9789264040014-en.

Ögren, M., Nyström, M., & Jarodzka, H. (2017). There's more to the multimedia effect than meets the eye: is seeing pictures believing? *Instructional Science*, 45(2), 263–287. https://doi.org/10.1007/s11251-016-9397-6.

Ott, N., Brünken, R., Vogel, M., & Malone, S. (2018). Multiple symbolic representations: the combination of formula and text supports problem solving in the mathematical field of propositional logic. *Learning and Instruction*, 58(December 2018), 88–105. https://doi.org/10.1016/j.learninstruc.2018.04.010.

Paivio, A. (1986). *Mental representations*. New York: Oxford University Press.

Pande, P., & Chandrasekharan, S. (2017). Representational competence: towards a distributed and embodied cognition account. *Studies in Science Education*, 53(1), 1–43. https://doi.org/10.1080/03057267.2017.1248627.

Park, B., Moreno, R., Seufert, T., & Brünken, R. (2011). Does cognitive load moderate the seductive details effect? A multimedia study. *Computers in Human Behavior*, 27(1), 5–10. https://doi.org/10.1016/j.chb.2010.05.006.

Ramjan, L. M. (2011). Contextualism adds realism: nursing students' perceptions of and performance in numeracy skills tests. *Nurse Education Today*, 31(8), e16–e21.

Reusser, K. (1996). From cognitive modeling to the design of pedagogical tools. In S. Vosniadou, E. De Corte, R. Glaser, & H. Mandl. (Eds.), *International perspectives on the design of technology supported learning environments* (pp. 81–104). Mahwah: Lawrence Erlbaum Associates, Publishers.

Rop, G. (2017). *Effects of task experience on attention to extraneous information during multimedia learning effects of task experience on attention* (Doctoral dissertation). Retrieved from https://doi.org/10.13140/RG.2.2.19515.31524

Rop, G., Verkoeijen, P. P. J. L., & van Gog, T. (2017). With task experience students learn to ignore the content, not just the location of irrelevant information. *Journal of Cognitive Psychology*, 29(5), 599–606. https://doi.org/10.1080/20445911.2017.1299154.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park: Sage.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: prevention, assessment, and adjustments. New York: Wiley.* https://doi.org/10.1002/0470870168.

Saß, S., & Schütte, K. (2016). Helping poor readers demonstrate their science competence: item characteristics supporting text-picture integration. *Journal of Psychoeducational Assessment*, 34(1), 91–96. https://doi.org/10.1177/0734282915588389.

Saß, S., Wittwer, J., Senkbeil, M., & Köller, O. (2012). Pictures in test items: effects on response time and response correctness. *Applied Cognitive Psychology*, 26(1), 70–81.

Saß, S., Schütte, K., & Lindner, M. A. (2017). Test-takers' eye movements: effects of integration aids and types of graphical representations. *Computers in Education*, 109, 85–97. https://doi.org/10.1016/j.compedu.2017.02.007.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, 48, 37–50. https://doi.org/10.1016/j.intell.2014.10.003.

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: behavior, theory, and biological mechanisms. *Cognitive Science*, 27(3), 525–559. https://doi.org/10.1016/S0364-0213(03)00011-9.

Schneider, S., Nebel, S., & Rey, G. D. (2016). Decorative pictures and emotional design in multimedia learning. *Learning and Instruction*, 44, 65–73. https://doi.org/10.1016/j.learninstruc.2016.03.002.

Schnotz, W. (2014). Integrated model of text and picture comprehension. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 72–103). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.006.

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, 13(2), 141–156. https://doi.org/10.1016/S0959-4752(02)00017-8.

Schüler, A., Pazzaglia, F., & Scheiter, K. (2019). Specifying the boundary conditions of the multimedia effect: the influence of content and its distribution between text and pictures. *British Journal of Psychology*, 110(1), 126–150.

Schwert, P. M. (2007). Using sentence and picture clues to solve verbal insight problems. *Creativity Research Journal*, 19(2–3), 293–306.

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory*, 18(7), 698–711. https://doi.org/10.1080/09658211.2010.506441.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6(1), 156–163. https://doi.org/10.1016/S0022-5371(67)80067-7.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing II: perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2), 127–189. https://doi.org/10.1037/0033-295X.84.2.127.

Solano-Flores, G., Wang, C., Kachchaf, R., Soltero-Gonzalez, L., & Nguyen-Le, K. (2014). Developing testing accommodations for English language learners: illustrations as visual supports for item accessibility. *Educational Assessment*, 19(4), 267–283.

Solano-Flores, G., Wang, C., & Shade, C. (2016). International semiotics: item difficulty and the complexity of science item illustrations in the PISA-2009 international test comparison. *International Journal of Testing*, 16(3), 205–219.

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology*, 54(10), 1046–1055. https://doi.org/10.1016/S0895-4356(01)00377-8.

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5.

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*. Lisse: Swets & Zeitling.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, 1(2), 112–125.

Whitley, K. N., Novick, L. R., & Fisher, D. (2006). Evidence in favor of visual representation for the dataflow paradigm: an experiment testing LabVIEW's comprehensibility. *International Journal of Human-Computer Studies*, 64(4), 281–303. https://doi.org/10.1016/j.ijhcs.2005.06.005.

Wiley, J., Sanchez, C. A., & Jaeger, A. J. (2014). The individual differences in working memory capacity principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 598–620). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.029.

Winter, P. C., Kopriva, R. J., Chen, C. S., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: results from a large-scale cognitive lab. *Learning and Individual Differences*, 16(4), 267–276. https://doi.org/10.1016/j.lindif.2007.01.001.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: implications for test development and measurement practice. *Applied Measurement in Education*, 22(2), 185–205. https://doi.org/10.1080/08957340902754650.

Wu, H. K., Kuo, C. Y., Jen, T. H., & Hsu, Y. S. (2015). What makes an item more difficult? Effects of modality and type of visual information in a computer-based assessment of scientific inquiry abilities. *Computers in Education*, 85, 35–48. https://doi.org/10.1016/j.compedu.2015.01.007.

Yang, D. C., & Huang, F. Y. (2004). Relationships among computational performance, pictorial representation, symbolic representation and number sense of sixth-grade students in Taiwan. *Educational Studies*, 30(4), 373–389. https://doi.org/10.1080/0305569042000310318.

Zahner, D., & Corter, J. E. (2010). The process of probability problem solving: use of external visual representations. *Mathematical Thinking and Learning*, 12(2), 177–204. https://doi.org/10.1080/10986061003654240.

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science*, 21(2), 179–217. https://doi.org/10.1207/s15516709cog2102_3.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, 18(1), 87–122.

Zhang, J., & Norman, D. A. (1995). A representational analysis of numeration systems. *Cognition*, 57(3), 271–295.

Zhang, J., & Patel, V. L. (2006). Distributed cognition, representation, and affordance. *Pragmatics and Cognition*, 14(2), 333–341. https://doi.org/10.1075/pc.14.2.12zha.

Zhao, F., Schnotz, W., Wagner, I., & Gaschler, R. (2014). Eye tracking indicators of reading approaches in text-picture comprehension. *Frontline Learning Research*, 6, 46–66. https://doi.org/10.14786/flr.v2i4.98.

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