Educational Psychology Review (2024) 36:49
https://doi.org/10.1007/s10648-024-09883-0

REVIEW ARTICLE

Revisiting Picture Functions in Multimedia Testing: A Systematic Narrative Review and Taxonomy Extension

Lauritz Schewior[1] · Marlit Annalena Lindner[1,2]

Accepted: 5 April 2024

Abstract

Studies have indicated that pictures in test items can impact item-solving performance, information processing (e.g., time on task) and metacognition as well as test-taking affect and motivation. The present review aims to better organize the existing and somewhat scattered research on multimedia effects in testing and problem solving while considering several potential moderators. We conducted a systematic literature search with liberal study inclusion criteria to cover the still young research field as broadly as possible. Due to the complexity and heterogeneity of the relevant studies, we present empirical findings in a narrative review style. Included studies were classified by four categories, coding the moderating function of the pictures investigated. The evaluation of 62 studies allowed for some tentative main conclusions: Decorative pictures did not appear to have a meaningful effect on test-taker performance, time on task, test-taking affect, and metacognition. Both representational and organizational pictures tended to increase performance. Representational pictures further seem to enhance test-taker enjoyment and response certainty. Regarding the contradictory effects of informational pictures on performance and time on task that we found across studies, more differentiated research is needed. Conclusions on other potential moderators at the item-level and test-taker level were often not possible due to the sparse data available. Future research should therefore increasingly incorporate potential moderators into experimental designs. Finally, we propose a simplification and extension of the functional picture taxonomy in multimedia testing, resulting in a simple hierarchical approach that incorporates several additional aspects for picture classification beyond its function.

Keywords Multimedia Effect · Testing · Assessment · Test Item Design · Problem Solving

Extended author information available on the last page of the article

Published online: 07 May 2024

Introduction

Students learn better from words and pictures than from words alone; this phenomenon is known as the multimedia effect in learning (Mayer, 2005). Multimedia design also plays an important role in educational testing materials. Visual representations are, for example, frequently used in large-scale international assessment materials such as the Program for International Student Assessment (*PISA*; OECD, 2009). Similar to multimedia learning (MML), studies have indicated that pictures in test items can increase item-solving performance (Hartmann, 2012; Lindner et al., 2017a, b; Ott et al., 2018; Saß et al., 2012) compared to text-only. This increase in test performance due to the use of pictures is also referred to as the *multimedia effect in testing* (Lindner, 2021; Lindner et al., 2017a, b).

However, research on multimedia testing (MMT) is still at an early stage. In contrast to multimedia learning (MML), where we can find a large number of reviews on the effects of multimedia design (Noetel et al., 2022), there is little systematic research on MMT, indicating the need for further investigation. Moreover, the term multimedia is strongly associated with learning research. As a result, MML and MMT are sometimes not clearly separated from each other. More work needs to be done to establish MMT as a conceptually and theoretically distinct line of research (e.g., Kirschner et al., 2017).

This review aims to contribute to this development by providing a systematic and comprehensive overview of the current state of research. Based on a systematic literature search, the review will address key issues in the field, such as an appropriate classification of picture types and their different effects on relevant educational outcomes, as well as identifying other potential moderator variables.

Multimedia Learning versus Multimedia Testing

Before outlining the aims of this review, key similarities and differences between MML and MMT should be highlighted due to their close relationship. In fact, the cognitive processes relevant to both MML and MMT overlap considerably. The learner ***and*** the test-taker must (at least in the initial phase of processing multimedia material) select information from the sensory memory (Atkinson & Shiffrin, 1971) and organize it in the limited working memory (Baddeley, 1992). Therefore, in both cases, it makes sense to explain observed cognitive facilitation with similar theoretical concepts, such as reduced cognitive load (Chandler & Sweller, 1991) due to using different cognitive channels (Clark & Paivio, 1991; Paivio, 1986). It is therefore reasonable that many authors in the field of testing (e.g., Dirkx et al., 2021; Hu et al., 2021; Jarodzka et al., 2015; Lindner, 2020) draw on multimedia learning theories that integrate these findings, such as the Cognitive Theory of Multimedia Learning (CTML; Mayer, 2005) and the Integrative Model of Text and Picture Comprehension (ITPC; Schnotz & Bannert, 2003). In contrast to the CTML, the ITPC more explicitly describes the integration processes and the different roles of verbal and visual elements and refers to a descriptive and a depictive branch of information processing. The model posits that text-based information (e.g., spatial relations) is often given only implicitly and requires interpretation, leading to the possibility of a text-based mental model that does not fully capture the relevant content. In contrast, pictorial information can be more directly extracted at a perceptual level. Thus, pictures can restrict the range of (mis)-interpretation derived from text, a concept known as the constraining interpretation function, as introduced by Ainsworth (2006). Thus, according to the ITPC, multimedia material (text + picture) facilitates the construction of a coherent mental model as text and pictures can have complementary functions (Schnotz & Bannert, 2003).

However, crucial differences between MML and MMT should be considered as well. One important difference lies in the structure of the presented material and the timing of its presentation. In contrast to learning situations, task-relevant material in testing situations is presented at the same time as the test question (Lindner, 2021). Thus, test takers always receive textual and pictorial information simultaneously and have the picture available during the *problem solving phase* (Lindner et al., 2017a, b), the characteristic cognitive phase of the problem-solving process (Novick & Bassok, 2005). Of course, assessments also take place in multimedia learning research and can include pictures as well (e.g., Scheiter & Eitel, 2015; Scheiter et al., 2014). However, these assessments are always directly related to a preceding multimedia learning phase. In fact, MML authors should note that using pictures in the learning phase *and* the subsequent assessment may be a factor to consider when evaluating learning outcomes (Lindner et al., 2021).

Another key difference is that both learning and testing contexts refer to fundamentally different goals. In learning, the central goal is that students acquire new knowledge and reproduce it after some time. A positive effect of multimedia on performance can thus be considered an inherent and desirable goal for both, students and educators. In testing, however, the primary goal of educators is not the improvement of student's performance, but rather the accurate measurement of it. In contrast to the learning context, it would be inappropriate to claim that higher performance due to the use of pictures is generally desirable in assessments. Whether, for example, performance-enhancing effects are considered "beneficial" or "positive" also depends on whether the targeted construct is measured more accurately or not (see, e.g., also Lindner, 2021).

Pictures can influence the validity of assessments in various ways, so understanding their effects is important. Keehner et al. (2023) for example, suggest using pictures in digital test items to support certain cognitive processes that are not critical for the measured target construct (e.g., reading in science items), in order to increase construct validity and reliability. Concerns about validity also arise in the context of low-stakes assessments, where there can be a significant disparity between an individual's actual proficiency and their demonstrated proficiency due to motivational constraints (Wise & DeMars, 2005). Knowledge about the motivational effects of pictures on students' test-taking effort could be relevant here, as there are indications that pictures can reduce rapid-guessing behavior in test situations (Lindner et al., 2017b).

Lastly, in both MML (Carney & Levin, 2002; Levin et al., 1987) and MMT (Hu et al., 2021; Lindner, 2021), authors have pointed out that pictures can have different functions, leading to different effects. However, there are differences in the description of these functions between MML and MMT. While five different functions, based on detailed exemplification (Levin, 1981), are described for MML (Carney & Levin, 2002*; decorational*, *representational*, *organizational*, *interpretational*; t*ransformational*), only four distinctive functions were derived from this classification and transferred to MMT (Elia & Philippou, 2004; Elia et al., 2007), which are described in the next section.

Picture Types in Multimedia Testing

Inspired by the prominent classification system from MML (Carney & Levin, 2002; Levin, 1981; Levin et al., 1987), four picture categories describing the relationship between a picture and the associated text have been proposed for MMT (Elia & Philippou, 2004; Elia et al., 2007). The definitions for *Decorative Pictures* (DPs) and *Representational Pictures* (RPs) are almost identical to those from MML. DPs can be loosely associated with the item context but do not display task or solution relevant information. RPs can represent a part or the whole of the (task-relevant) text content.

Regarding *Organizational Pictures* (OPs), there are differences between the definitions used in MML and MMT. In MML, it is emphasized that the function of OPs is to organize (complex) fragmented textual information more cohesively and thus to make the textual information easier to integrate during the learning process (Levin, 1981). The original definition refers to pictures that present text content sequentially (Carney & Levin, 2002), more similar to dynamic pictures (e.g., Ehrhart & Lindner, 2023) included in some studies of this review. In contrast, Elia and Philippou (2004) speak of OPs when they provide directions that support the solution procedure. More specifically, Lindner (2021) suggested that the term may be used especially for pictures that support the solution procedure by providing a visuospatial structure.

The category of *Informational Pictures* (IPs) has been explicitly coined for MMT. IPs depict information that is essential for the solution of the problem (Elia & Philippou, 2004; Elia et al., 2007). Consequently, a test taker must process IPs before answering the test questions. For IPs, there may also be cases where a picture is the only source of information for a given task, with no additional text provided (picture-only). Figure 1 provides an overview of the defining characteristics of different picture functions and includes examples from the MMT literature for each of the four categories.

This functional taxonomy was used in previous reviews of multimedia testing (see also Hu et al., 2021; Lindner, 2021) and seems to be the most appropriate for our endeavour as well. It provides the most explicit and clear-cut distinctions between picture categories as compared to other taxonomies. The function of a picture in terms of whether it is only loosely connected to the task text (i.e., decorative), double-codes task-relevant information from the text (i.e., representational, organizational), or complements the text with task-critical information (i.e., informational), is very clear to define and determine.

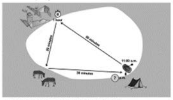
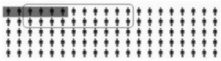
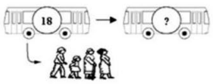
Type	Function/ Characteristics	Example from the Literature	Revised Taxonomy*
Decorative	**Decorative Pictures (DPs)** can be loosely associated with the item context, but do not have a relation to the task.	Anton has two identical toy buses and two equally sized rod magnets. He sets up the buses opposite each other and places one of the two rod magnets on each of the buses. The south pole of the magnet on one bus points to the south pole of the magnet on the other bus. Anton pushes the buses close together and then lets them go. What happens to the toy buses when Anton lets them go? A The buses attract each other. B The buses attract each other first and then repulse each other. C The buses are repulsing each other. D Nothing, the buses remain in place. **Lindner (2020)**	**Decorative Pictures** (Differentiation between *unrelated* vs. *contextual* DPs)
Representational	**Representational Pictures (RPs)** represent (i.e., double code) task-relevant information that is also given in a verbal format in a corresponding text.	Liam and his dad go camping near Hagen Lake. They want to visit some castle ruins on the other side of the lake. The two take a paddle boat and start their trip at 11:00 a.m. To get to the castle ruins, they need 50 minutes. There they take a one-hour break. After the break, they go on paddling for 20 minutes and discover some donkeys in an animal park near the shore. Then, they paddle back to the campsite. To cover this distance, they need 30 minutes.  1 hour 50 minutes 20 minutes 30 minutes 11:00 a.m. ? p.m. At what time do they reach the campsite? ○ They reach the campsite at 12:40 p.m. ○ They reach the campsite at 1:00 p.m. ○ They reach the campsite at 1:30 p.m. ○ They reach the campsite at 1:40 p.m. **Ehrhart & Lindner (2023)**	**Representational Pictures** (Classification of OPs as a subcategory of RPs)
Organizational	**Organizational Pictures (OPs)** represent aspects of the problem that is verbally described in a visual-spatial manner to support the solution process.	There is a newly discovered disease, Disease X, which is transmitted by a bacterial infection. Here is some information about the current research on Disease X and efforts to test for the infection that causes it. A person has 6 chances out of 100 of having the infection. There is a test to detect whether or not a person has this infection, but it is not perfect. Specifically, only 4 of the 6 chances of having the infection were associated with a positive reaction from the test. On the other hand, 16 of the remaining 94 chances of not having the infection (that is, being perfectly healthy) were also associated with a positive reaction from the test. The picture below summarizes the above information, and is provided for you to use in the process of answering the following item. All the chances in this situation are represented by the largest circle. The smallest circle represents the chances of having the infection. The slightly larger, overlapping circle represents the chances of a positive reaction to the test.  Imagine Michael is tested now. Out of a total of 100 chances, Michael has _____chance(s) of positive reaction from the test, ____ of which will be associated with actually having the infection. **Brase (2009)**	
Informational	Informational Pictures (IPs) provide additional information which is not given in the text, but is essential for solving the task.	There were 18 people on the bus. Some people got out. How many people are now on the bus?  Answer = ... **Berends & van Lieshout (2009)**	**<u>Informative</u> Pictures** (Category renamed to be more intuitively understandable)

Fig. 1 Key Characteristics of Picture Function Types and Examples for a Decorative (Lindner, 2020), Representational (Ehrhart & Lindner, 2023), Organizational (Brase, 2009) and Informational (Berends & van Lieshout, 2009) Picture. Pictures are Adapted from Originals in Size. **Note.** Reprinted with permission from [1]Elsevier and [3]John Wiley and Son or under [2]CC BY license. *The revised and simplified taxonomy is a result of this narrative review and will be explained in greater detail in the Discussion section and is displayed in Fig. 6 in its hierarchical structure

However, more taxonomies exist in the MML literature to classify pictures in multimedia learning contexts. Ainsworth (1999, 2006) also took a functional approach to describe a taxonomy for multiple representations. The functions described in her classification overlap partly with the picture functions we use and propose in this review for the context of multimedia testing. For example, for an item with IPs, one could argue that the text and the picture have complementary functions (Ainsworth, 1999). RPs, on the other hand, are particularly well suited to fulfil a constraining interpretation function (Ainsworth, 1999).

Some other taxonomies focus on characteristics of pictures that do not refer to the picture function (cf. Slough & McTigue, 2013): Alesandrini (1984), for example, classified pictures according to their visual similarity to the relevant object or concept, distinguishing between representational, analogical, and arbitrary representations. Hegarty et al. (1991) focused on the level of abstraction and distinguish between iconic and schematic representations. In fact, pictures in the same functional category can still vary considerably also in the testing context, for example, in their level of abstraction (for iconic RPs see, e.g., Lindner, 2020; for abstract RPs see, e.g., Ott et al., 2018). Accordingly, a multidimensional classification system of pictures is imaginable (e.g., function, abstractness, dynamism, complexity). However, a more detailed categorization would not improve the present review as the required information regarding the picture material is not reported in most primary studies so far. Our systematic overview of the field is therefore structured on the basis of the four established picture functions in MMT (Elia & Philippou, 2004; Hu et al., 2021; Lindner, 2021). Nevertheless, while building on previous work in multimedia testing, our goal is to further discuss and expand the existing taxonomy and make suggestions for refinements and future directions in the context of this review, as we lay out in the Discussion part of this article.

Moderators Beyond Picture Types

Other potential moderators of picture effects in testing may play a role as well. At the item-level, variables worth considering could be the task and item response format (Lindner et al., 2022), the item complexity (Hoogland et al., 2018b) or the subject domain (Lindner, 2020). At the test-taker level, cognitive (for domain specific abilities see, e.g., Berends & van Lieshout, 2009; for prior knowledge see, e.g., Ehrhart & Lindner, 2023) and affective-motivational moderators (e.g., students' domain-specific attitude; Cooper et al., 2018) could be considered. Additionally, examinees' age may play a part in the effects. A comprehensive and systematic investigation of different moderating variables is still required in the field, which is one important goal of this review.

Moreover, studies concerned with multimedia effects in testing come from a wide range of research areas. Many studies were established in educational contexts and focused on mathematics (e.g., Cooper et al., 2018; Ehrhart & Lindner, 2023) and science problems (e.g., Lindner et al., 2018). Other studies are focusing on applied contexts, such as medical diagnostics (e.g., Garcia-Retamero et al., 2010) or programming tasks (Whitley et al., 2006). As a result, the rather small research base is also very heterogeneous with substantial differences in the tasks and domains in which investigations have taken place. This makes it even more difficult to compare and interpret the results and demands further investigation.

Typical Outcome Measures

Outcome measures in multimedia testing can be categorized in three broad categories: Cognitive, metacognitive, and affective-motivational outcomes. *Cognitive* outcomes that are typically measured are performance (e.g., Saß et al., 2012) and time on task (e.g., Berends & van Lieshout, 2009). Whereas test accuracy is a key indicator of test-taker performance, time on task is a parameter for ongoing cognitive processes in multimedia testing. Few studies have used other measures of mental effort, for example self-reports (Wang et al., 2022), secondary task score (e.g., Dindar et al., 2015) or eye-tracking data (e.g., Zheng & Cook, 2012). *Metacognitive* outcomes reflect students' cognition regarding their own performance in the test situation, such as the self-perceived certainty of their response correctness (Lindner et al., 2021). *Affective-motivational* outcomes comprise affective experiences such as test anxiety but also motivational outcomes such as test taking motivation (Lindner et al., 2021). However, there is still no systematic overview of how often the different groups of outcome measures have been studied in MMT and how heterogeneous the operationalisation had been.

The Present Review

In a first meta-analysis, Hu et al. (2021) refer to the function of a picture as a central moderator of multimedia effects in the test context. While they found small-to-medium multimedia effects on test performance for RPs (Hedges's $g = 0.24$) and OPs (Hedges's $g = 0.52$), they found no significant summary effect for IPs. However, given their search period (until August 2018), Hu et al. (2021) only identified 26 studies and were, for example, not able to investigate the effects of DPs in testing situations. A complementary conceptual approach to discuss the base and recent developments in the field of MMT is provided by Lindner (2021). While this work is a narrative reflection on the field, it does not explicitly feature a systematic search and analysis of the literature. The aim of the present review is to systematically investigate—for the first time—the effects all four types of picture functions. Another contribution to the existing work concerns the classification of pictures in MMT. While Hu et al. (2021) and Lindner (2021) also refer to the functional classification approach, we systematically investigate how frequently the respective terminology is actually used in primary studies. In the present work, we also aim to discuss simplifications and potential improvements of the classification system for pictures in MMT to motivate the field towards using a more unified terminology in future studies.

Another contribution of the present review is that it has an additional focus on other potential moderators, e.g., at the item level or at the task-taker level. A systematic literature overview of moderators that have been investigated in primary studies does not yet exist. Again, broader inclusion criteria are needed here, as some studies do not have a control group because they are investigating a potential moderator only (e.g., Dewolf et al., 2017; Dirkx et al., 2021; Saß et al., 2017). To achieve this goal, we used all available studies for this investigation, not only studies that would meet strict statistical and study design criteria for a meta-analysis.

Our narrative review is primarily exploratory in its motivation to capture the field in its breadth, focusing on the effects of different picture types and reflecting on potential moderating factors. Still, based on the literature, it was possible to formulate hypotheses regarding the expected effects of specific picture functions on certain outcomes. For DPs, for example, a multimedia effect seemed unlikely, neither with regard to the cognitive and metacognitive variables or the affective-motivational outcomes based on large studies (Ehrhart et al., 2024; Lindner, 2020). On the contrary, we expected that individuals show increased test-taking performance when solving text-picture items with visualizations that can be classified as representational or organizational (e.g., Hu et al., 2021). Moreover, it can be expected that positive effects of RPs on metacognitive and affective-motivational variables are evident as well (Lindner, 2020; Lindner et al., 2022). For IPs, the data base seemed to be complex and heterogeneous (Hu et al., 2021), which made it more difficult to form specific hypotheses. Yet, we expected to find the heterogeneity of IP effects reflected in our results.

Taken together, the present work is, by implementing very broad inclusion criteria, the first systematic review that gives a comprehensive overview of the existing literature on multimedia testing. In addition to providing a numerical overview of the number of studies and outcome measures that have been examined so far, the narrative format offers the opportunity to qualitatively summarize the very heterogeneous field, which is relevant in the development towards a more distinctive line of research.

Methods

Search Method and Inclusion Criteria

The ScienceDirect, PsycINFO and Web of Science databases were searched with the following search string: “illustration” OR “Multimedia effect” OR “Multiple representation” OR “text-picture” OR “external representation”) AND (“problem solving” OR “test” OR “testing” OR “assessment”. No limit in terms of older publications has been defined, whereas studies until December 2022 were included in the review. The search string was chosen after carefully analyzing keywords from prominent studies in the field. Due to the ambiguous terminology across articles and fields, we assumed that several articles that would be of interest for the review were not found by searching the databases. Therefore, the snowball method was another vital component of the search strategy. “Snowballing” involves tracking references of already identified articles, which can significantly contribute to the number of studies found (Greenhalgh & Peacock, 2005). The database was further extended by studies that we had already collected from the field in several years of research. These studies were also reviewed in more detail regarding their suitability for this work.

The essential criterion for including articles was that pictures are displayed in a test or problem-solving material where the text-based problem statement or question is displayed simultaneously with the (manipulated) picture information. A control group (e.g., text-only) was considered as beneficial but was not mandatory for inclusion in this narrative review. Studies that have been published in English and

German language were included. The review incorporates not only articles from scientific journals but also two dissertations and one research report.

Screening Procedure

After the database search, a detailed screening of all results was performed. The screening consisted of three parts: First, only the titles of all results from the databases were read. Here, only studies were excluded that clearly belonged to an irrelevant and unrelated research area. A conservative approach was taken in order to avoid excluding potentially relevant studies at this stage.

Second, for the remaining articles, the abstracts were compiled and screened. In this phase, studies found via the snowball method were additionally screened. The leading single reason for exclusion at this stage was that the study in question came from the field of MML and not MMT.

Third, the full texts were collected and analyzed in detail for all studies that appeared to be appropriate for the review according to the information in the abstracts. In this phase of a more detailed examination of the full texts, the most common reason for exclusion was still that the decisive criterion of simultaneous presentation of pictures and tasks in multimedia testing was often not met (i.e., MML vs. MMT). Figure 2 illustrates the entire screening procedure in a PRISMA flow diagram.

Data Extraction and Coding

Data extraction was carried out systematically according to a predefined coding table. The resulting excel sheet was the basis for the systematic analysis of the results and the narrative writing process. Information that had been coded was: Authors; Year; Journal; Picture type (with sample image); Outcome variables; Moderator variables; Domain; Item description; Number of items; Answer format; Context (e.g., school or university); Testing technique (paper–pencil vs. computer-based); Number of participants; Mean age of participants; Country; Main results; Other (e.g., use of Eye-Tracking).

Following our narrative approach, we only report broader categories of “positive”, “neutral” or “negative” effects and do not apply statistical tests. For the interpretation of our findings, it is therefore important to note that the interpretation of whether the increased or decreased performance due to the use of pictures can actually be considered as a “positive” or as “negative” outcome in practice does primarily depend on the pictures’ impact on the test validity and the specific goal of a given assessment. Thus, the terms “positive” and “negative” in the results section should only be interpreted in terms of an effect direction.

Results and Narrative Synthesis

Investigated Picture Types and Applied Terminology Sixty-two articles were coded for this review which were classified by the investigated picture categories. The coding of the picture functions was performed by two coders and resulted in a very high

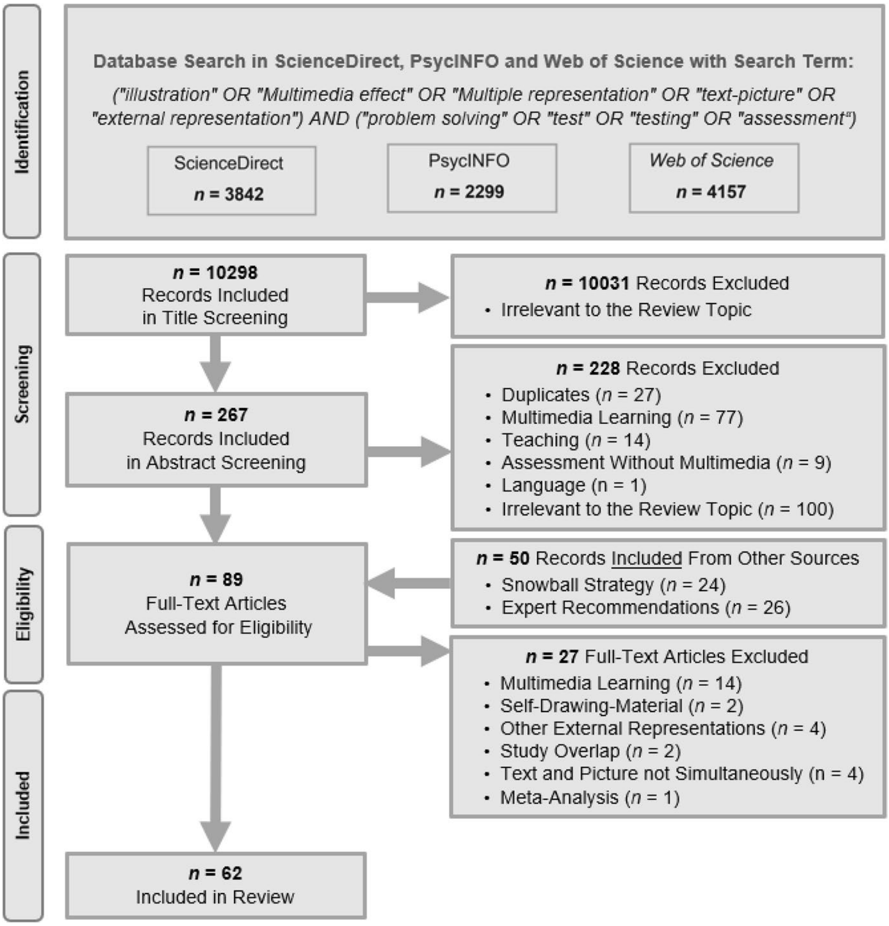

Identification
Database Search in ScienceDirect, PsycINFO and Web of Science with Search Term:
("illustration" OR "Multimedia effect" OR "Multiple representation" OR "text-picture" OR "external representation") AND ("problem solving" OR "test" OR "testing" OR "assessment")
ScienceDirect n = 3842
PsycINFO n = 2299
Web of Science n = 4157
Screening
n = 10298 Records Included in Title Screening
n = 10031 Records Excluded
• Irrelevant to the Review Topic
n = 267 Records Included in Abstract Screening
n = 228 Records Excluded
• Duplicates (n = 27)
• Multimedia Learning (n = 77)
• Teaching (n = 14)
• Assessment Without Multimedia (n = 9)
• Language (n = 1)
• Irrelevant to the Review Topic (n = 100)
Eligibility
n = 89 Full-Text Articles Assessed for Eligibility
n = 50 Records Included From Other Sources
• Snowball Strategy (n = 24)
• Expert Recommendations (n = 26)
n = 27 Full-Text Articles Excluded
• Multimedia Learning (n = 14)
• Self-Drawing-Material (n = 2)
• Other External Representations (n = 4)
• Study Overlap (n = 2)
• Text and Picture not Simultaneously (n = 4)
• Meta-Analysis (n = 1)
Included
n = 62 Included in Review

Fig. 2 Flow Diagram of Article Search and Screening

inter-rater reliability (Cohen's Kappa = 0.957). For two articles, there was an initial disagreement among the raters due to apparent translation issues of the items, which could be resolved by discussion in both cases. In twelve studies, more than one picture type was investigated in separate experimental conditions. Thus, a study may appear in the result sections of more than one function type (e.g., Berends & van Lieshout, 2009; Ehrhart & Lindner, 2023; Lindner, 2020).

Overall, pictures that can be classified as representational pictures (RPs) were examined most frequently (35.71%), followed by informational pictures (IPs; 32.86%), decorative pictures (DPs; 20.0%) and organizational pictures (OPs; 11.43%). However, we found that not all studies applied the proposed terminology for the respective picture functions. While the term "decorative" was used in 50% of studies that investigated pictures that can be characterized as DPs, this proportion was slightly higher for RPs (56%), but substantially lower for OPs (12.5%) and IPs

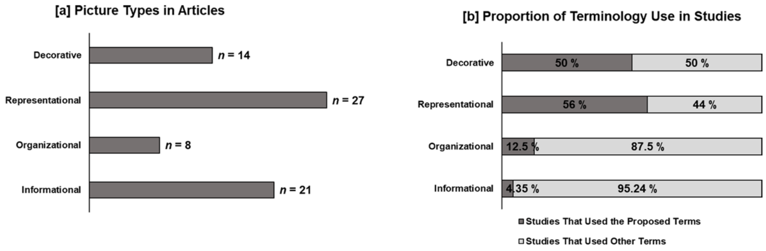


Fig. 3 (**a**) Number of Articles (n) Investigating a Certain Picture Type. (**b**) Percentage of Articles that Used the Proposed Terminology to Describe the Function of the Investigated Picture Type as Identified in this Review

Table 1 Terms Used by Different Authors for Different Types of Pictures

Picture Function	Other Terms
Decorative	*useless* (Berends & van Lieshout, 2009); *vignette illustrations* (Solano-Flores et al., 2014); *contextual* (Clinton & Walkington, 2019); *misleading* (Clinton & Walkington, 2019); *irrelevant* (Clinton & Walkington, 2019); *illustrations* (Cooper et al., 2018); *seductive picture* (Strobel et al., 2019)
Representational	*helpful illustrations* (Berends & van Lieshout, 2009); *information equivalent representation* (Saß & Schütte, 2016); *redundant text-picture combination* (Magnus et al., 2020); *repeated illustrations* (Wang et al., 2022); *non-interactive graphic representation* (Zheng & Cook, 2012); *graphics* (Dindar et al., 2015; Malone et al., 2020; Ott et al., 2018); *graphs* (Ögren et al., 2017); *diagram + illustration* (Cooper et al., 2018); *diagrammatic illustrations* (Clinton & Walkington, 2019); *diagram* (Beveridge & Parkins, 1987)
Organizational	*venn diagram* (Brase, 2009); *dotted venn diagram* (Brase, 2009); *icon representation* (Brase, 2009); *visual aid* (Garcia-Retamero & Hoffrage, 2013); *icon arrays* (Garcia-Retamero et al., 2010); *diagram* (Booth & Koedinger, 2012; Cooper et al., 2018); *external representation* (Múñez et al., 2013); *illustrations* (Chu et al., 2017)
Informational	*essential* (Berends & van Lieshout, 2009); *complementary* (Saß & Schütte, 2016); *required pictures* (Wang et al., 2022); *non-redundant text–picture combination* (Magnus et al., 2020); *pictorial representation* (Lin et al., 2013; Yang & Huang, 2004); *close to real-life representations* (Hoogland et al., 2018a, b); *visualisation* (Dirkx et al., 2021); *contextualised* (Ramjan, 2011); *content illustration* (Wu et al., 2015); *images* (Vorstenbosch et al., 2013); *pictorial information* (Jarodzka et al., 2015); *visual programming language* (Whitley et al., 2006); *pictures* (Garret, 2008; Goolkasian, 1996, 2000; Goolkasian, 1996, 2000; Hartmann, 2012; Saß et al., 2012; Schnotz & Wagner, 2018; Schnotz et al., 2014); *anatomical illustration* (Bahlmann, 2018)

(4.35%). Thus, although studies have investigated pictures that meet the definitions of OPs and IPs, they rarely referred to them as such. Figure 3 provides an overview of how many articles studied the different types of pictures, along with the percentage of studies that employed the proposed labels for the picture functions.

Furthermore, in Table 1, we have collected terms used by authors in the primary studies that differ from the functional terminology we have used. It illustrates that the functional approach terminology is not used consistently in the field. However, Table 1 also demonstrates that many authors have implicitly employed a functional approach when describing their pictures. They referred, for example, to decorative pictures as “useless” or “irrelevant”, to representational pictures as “helpful” or “information-equivalent representation”, and to informational pictures as “essential” or “required pictures”. Furthermore, Table 1 shows that no alternative taxonomy seems to have gained widespread acceptance in the field of MMT yet, as there is no discernible systematic pattern in the used terms in primary studies.

Seven studies (Ahmed et al., 2021; Crisp & Sweiry, 2006; Fırat, 2017; Gray et al., 2012; Günbaş, 2020; Hao, 2010; Solano-Flores et al., 2016) could not be grouped in one of the categories because they presented different types of pictures in one experimental condition.

Identified Outcome Measures in Primary Studies For all picture types, performance is the most frequently studied outcome. For OPs and IPs in particular, only a small proportion of studies have collected data on other relevant outcomes. The most comprehensive database in terms of a variety of outcome measures is available for RPs. For RPs, a notable number of seven studies have also collected eye-tracking data (Lindner et al., 2017a, b; Malone et al., 2020; Ögren et al., 2017; Ott et al., 2018; Saß et al., 2017; Wang et al., 2022; Zheng & Cook, 2012). The results of the eye-tracking studies are mentioned and integrated into the narrative synthesis where appropriate and meaningful. It should also be mentioned that some studies reported qualitative measures (Ahmed et al., 2021; Gray et al., 2012; Hao, 2010). However, these variables are not part of the present review. Figure 4 shows how often the different outcome measures were investigated in primary studies, categorized by the four distinct picture types.

Multimedia Effects on Performance There are slight differences in the dependent variables related to performance. Most studies reported the proportion of correct responses

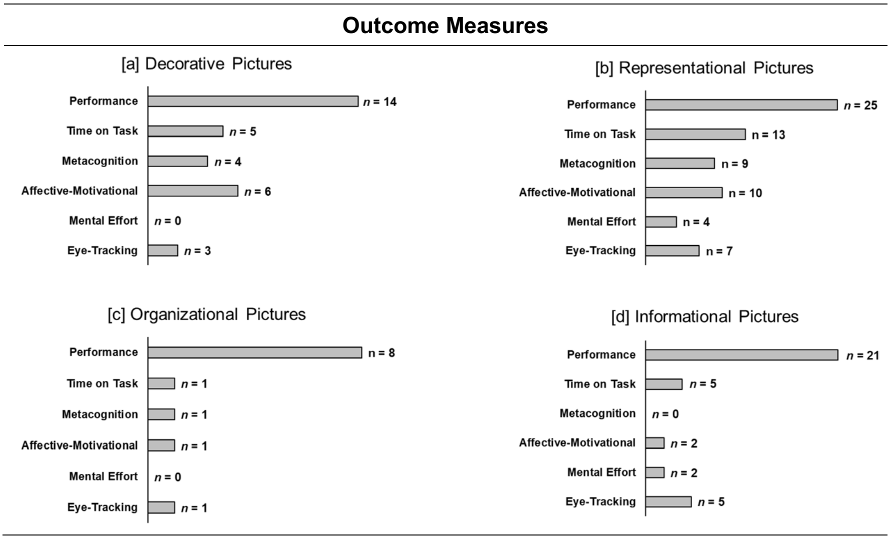


Fig. 4 Number of Articles (n) in Which the Various Outcome Measures were Investigated for (**a**) Decorative, (**b**) Representational, (**c**) Organizational and (**d**) Informational Pictures

given by test-takers. However, some studies investigated the proportion of responses that were realistic (Dewolf et al., 2014, 2015, 2017), and some studies reported differences between text-only and text-picture items in terms of item difficulty (Bahlmann, 2018; Dirkx et al., 2021; Lindner et al., 2018; Vorstenbosch et al., 2013).

For *DPs*, twelve out of fourteen studies compared a text-picture condition with a text-only condition. None of these twelve studies found a significant effect of DPs on performance compared to text-only (Agathangelou et al., 2008; Berends & van Lieshout, 2009; Clinton & Walkington, 2019; Cooper et al., 2018; Dewolf et al., 2014, 2015; Ehrhart et al., 2024; Elia et al., 2007; Lindner, 2020; Solano-Flores et al., 2014; Strobel et al., 2019). Thus, DPs seemed to neither increase nor decrease performance. This finding is in line with results from research on MML (Carney & Levin, 2002; Levin et al., 1987). Our results suggest that DPs in MMT demand little test-taker attention (Lenzner et al., 2013) rather than reducing performance as seductive details (Lehman et al., 2007).

Regarding *RPs*, conclusions about the effects on performance compared to text-only could be drawn from twenty-one articles. An apparent majority of articles, namely fourteen, reported higher accuracy rates for items with RPs compared to text-only (Beveridge & Parkins, 1987; Cooper et al., 2018; Ehrhart & Lindner, 2023; Ehrhart et al., 2024; Lindner, 2020; Lindner et al., 2018, 2021, 2022, 2017a, b; Malone et al., 2020; Saß et al., 2012; Wang et al., 2022). Six studies found no effects (Agathangelou et al., 2008; Berends & van Lieshout, 2009; Clinton & Walkington, 2019; Ögren et al., 2017; Ott et al., 2018; Zheng & Cook, 2012) and in one study it was reported, that effects were dependent on a moderator (Magnus et al., 2020). Taken together, the evaluated studies with RPs show clear evidence for a multimedia effect in testing that is consistent with the multimedia effect in learning (Levin et al., 1987; Mayer, 2005).

For *OPs*, a similar picture emerges from the identified studies. From seven articles with a text-only control group, six reported higher scores for items with OPs compared to text-only (Brase, 2009; Chu et al., 2017; Cooper et al., 2018; Garcia-Retamero & Hoffrage, 2013; Garcia-Retamero et al., 2010; Múñez et al., 2013). One study reported a higher accuracy rate as a function of a moderator variable, namely the age of the participants (Booth & Koedinger, 2012).

The effects of *IPs* on performance compared to text-only could be derived from sixteen articles. Six articles found that items with IPs increased the performance compared to text-only (Goolkasian, 1996; Hartmann, 2012; Hoogland et al., 2018b; Ramjan, 2011; Saß et al., 2012; Whitley et al., 2006). No effects of IPs on performance compared to text-only were shown in three studies (Bahlmann, 2018; Garret, 2008; Goolkasian, 2000). Unlike the other three categories, five articles with IPs found a decreased performance (Berends & van Lieshout, 2009; Elia et al., 2007; Hunt, 1978; Lin et al., 2013; Yang & Huang, 2004). In two articles, no clear conclusion was reported. Instead, it was noted that the type of task seems to be a key moderator (Magnus et al., 2020; Vorstenbosch et al., 2013).

While we have drawn relatively straightforward conclusions regarding the effects of DPs, RPs and OPs on performance, this is not possible for IPs, given the present database. Remarkably, IPs are the only category for which performance-reducing effects were reported. Previous reviews have already indicated the strong heterogeneity in studies that investigate IPs (Lindner, 2021) and this is, once again, confirmed in our review for a broader data base. Figure 5 illustrates the reported directions of multimedia effects compared to text-only for all four outcome measures and categorized by the four distinct picture types.

Multimedia Effects on Time on Task Five studies assessed the effects of *DPs* on time on task. Whereas Berends and van Lieshout (2009) found that math items without pictures were processed significantly faster than tasks with DPs, three studies found no significant effects of DPs on time on task (Ehrhart & Lindner, 2023; Ehrhart et al., 2024; Lindner, 2020). In an eye-tracking study, a longer time on task was found in the experimental condition with DPs (Strobel et al., 2019). Findings from this and other eye-tracking studies showed, however, that DPs are rarely (Dewolf et al., 2015) and shortly (Strobel et al., 2019; Wang et al., 2022) fixated by test takers. Overall, the assumption that DPs have little effect on the cognitive processes of test takers seems to be supported here.

Regarding *RPs*, most articles found no effect on time on task compared to text-only (Ehrhart & Lindner, 2023; Ehrhart et al., 2024; Lindner, 2020; Lindner et al., 2017a; Malone et al., 2020; Ögren et al., 2017; Ott et al., 2018; Zheng & Cook,

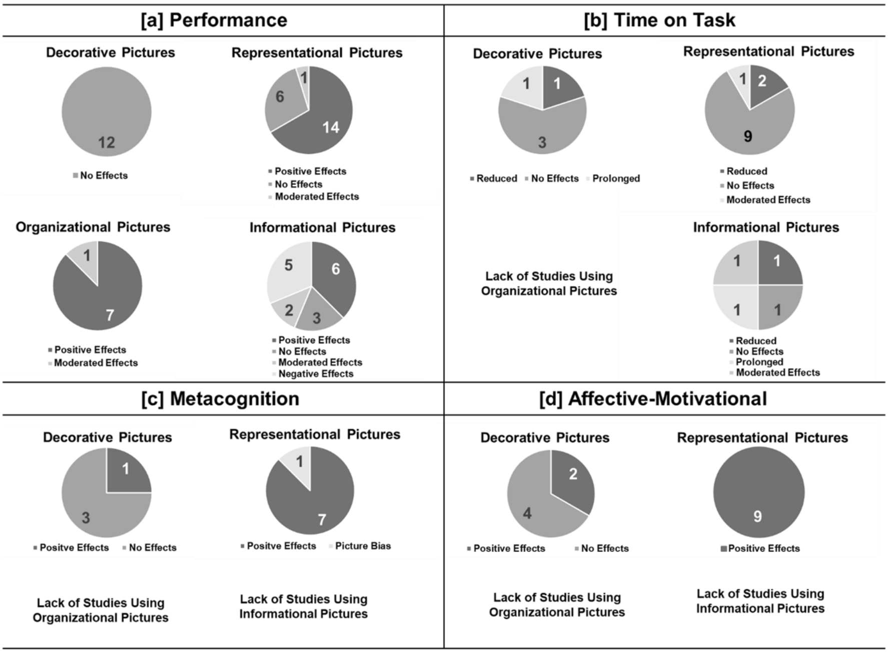


Fig. 5 Direction of Reported Picture Effects on (**a**) Performance, (**b**) Time On Task, (**c**) Metacognition and (**d**) Affective-Motivational Outcomes Compared to Text-Only for Studies with Decorative, Representational, Organizational and Informational Pictures. **Note.** The terms positive and negative are used in this context to refer to the direction of the statistical effects. A positive effect on performance can be understood as a performance-enhancing effect. However, it highly depends on the goals of the test and the test designer whether this would be considered a beneficial outcome in a particular assessment

2012). On the other hand, Lindner et al., (2017b) and Saß et al. (2012) even found a reduction in time on task due to RPs whereas Berends and van Lieshout (2009) reported a prolonged time on task. Although several studies do not show an impact on the total time on task, eye-tracking studies suggest an effect on cognitive processes. Students looked less often at text-based parts of the item, such as the text stem or the problem statement (Lindner et al., 2017a; Ögren et al., 2017). Thus, students seem to use the time they would usually spend on textual information in text-only items to process the pictorial information in text-picture items. Moreover, RPs seem to make cognitive processes more efficient during the initial problem representation phase (i.e. mental scaffolding) and were also attended to for mental model updating in the later problem-solving phase (Lindner et al., 2017a).

For *OPs*, only Múñez et al. (2013) found that time on task was reduced compared to text-only items. An eye-tracking study with OPs and RPs (but without a text-only control group) reported that the relative fixation duration on the picture was higher for items with OPs compared to items with RPs (Saß et al., 2017).

Regarding *IPs*, similar to results on performance, contrasting effects on time on task are reported. Studies found a prolonged time on task (Berends & van Lieshout, 2009), a reduced time on task (Goolkasian, 1996) or no effect (Saß et al., 2012). In one study (Whitley et al., 2006), the time on task of test-takers depended on the specific type of task. Consistent with the definition of IPs (they present additional information), an eye-tracking study by Wang et al. (2022) found that fixation transitions between text and pictures were more frequent in items with IPs compared to items with DPs or RPs.

Multimedia Effects on Metacognition For *DPs,* three studies found no effects on perceived item ease (Lindner, 2020) or on a metacognitive expectation of the correctness of responses (e.g., Ehrhart et al., 2024), whereas the results of one study indicate that DPs can increase the response certainty compared to text-only (Dewolf et al., 2015). Although the data base is very small, the results appear to be consistent with a lack of effects on cognitive measures.

Regarding *RPs,* seven articles reported significant effects on metacognition. Among them were a reduction in perceived item ease (Lindner, 2020), lower item-difficulty ratings (Malone et al., 2020) and higher metacognitive expectation of the correctness of responses (Ehrhart & Lindner, 2023; Ehrhart et al., 2024; Lindner et al., 2021, 2022). Ögren et al. (2017) found a "picture bias" in their study. Students were more likely to agree on a true or false task when RPs were present. Thus, identified studies indicate that RPs can significantly affect test-takers' metacognitive processes and evaluations. However, the metacognitive outcomes measured were manifold and the articles in the database referred to several slightly different variables, which makes it harder to compare them across studies.

For *OPs*, we found only two relevant studies. Garcia-Retamero and Hoffrage (2013), reported on a task-difficulty rating and found a main effect of the presentation format (task with pictures were rated as less difficult). Chu et al. (2017) found that students made fewer conceptual errors in text-picture items compared to text-only items.

For *IPs*, we were not able to identify any studies that have investigated effects on metacognition.

Multimedia Effects on Affective-Motivational Outcomes For *DPs,* four articles reported no effects on affective-motivational outcomes, namely students' interest (Clinton & Walkington, 2019) and item-solving satisfaction (e.g., Ehrhart et al., 2024; Lindner, 2020). In contrast, Cooper et al. (2018) reported that students rated tasks more favourably when decorative elements were presented. Agathangelou et al. (2008) also speak of a positive attitude towards DPs.

For *RPs*, nine studies found significant effects of RPs on affective-motivational outcomes. For example, RPs positively affected item-solving satisfaction (Ehrhart & Lindner, 2023; Ehrhart et al., 2024; Lindner, 2020; Lindner et al., 2018, 2022) or significantly reduced the average level of rapid-guessing behaviour (as an indicator for test-taking engagement; Lindner et al., 2019, 2017a, b). Two studies reported a positive attitude of students' toward RPs (Agathangelou et al., 2008; Malone et al., 2020). Thus, there are clear indications that RPs increase the enjoyment and motivation of test participation. Although similar findings may be expected for *OPs*, we can only report the lack of respective studies.

For *IPs*, Hunt (1978) found that students perceived tests with pictures as a more valid measurement of their competencies compared to text-only tests. Ramjan (2011) reported that students valued tasks with IPs more highly and rated the tasks as more relevant.

Moderators: Test-Taker Level

Student Abilities / Prior Knowledge The category most frequently researched at the test-taker level pertains to the test-takers' capabilities. A slightly greater proportion of studies in our database indicate that the level of student abilities is not a significant moderator of multimedia effects in testing. Studies reported no significant moderating effect of arithmetical abilities (Berends & van Lieshout, 2009 for DPs, RPs and IPs; Chu et al., 2017 for OPs), cognitive abilities (Lindner et al., 2018 for RPs; Múñez et al., 2013 for OPs; Strobel et al., 2019 for DPs) or reading abilities (Lindner et al., 2018 for RPs). However, other findings suggest a certain moderating influence: Students from lower-track schools, who, on average, have lower cognitive skills than students from higher-track schools, were particularly affected from (representational) pictures in two studies (Ehrhart & Lindner, 2023; Lindner et al., 2022). Regarding the influence of prior knowledge, Ehrhart and Lindner (2023) suggested a stronger effect of RPs for students with low to medium prior knowledge. In contrast, Cooper et al. (2018) indicated that students with higher mathematical skills could use RPs more efficiently. However, the lack of data and the heterogeneity in operationalizations make a sound interpretation difficult. More primary studies are needed for conclusions and standardized assessments of cognitive abilities and prior knowledge should be used to improve the comparability of results.

Age Regarding the age, Booth and Koedinger (2012) found that students from the seventh and the eight-grade showed higher scores in algebra items with RPs, whereas students from the sixth grade performed better in the text-only condition. Lindner et al. (2018) also reported that age was a significant predictor in their model. However, the data indicate that younger students (mostly 5th grade) benefited more from the presence of RPs in science items then older students (mostly 6th grade). Even though we lack clear evidence in this regard, an interaction of students' age and picture type may exist, depending on the prerequisite knowledge needed to interpret the picture properly. The more abstract or complex the picture and training needed for its correct interpretation (e.g., algebra), the more may students with a higher age and higher level of education benefit from the picture, whereas pictures that are realistic depictions and do not need training for interpretation may better support students with lower age and status of education. However, this is only an initial assumption that needs to be tested in future research.

Affective-Motivational Factors We have already covered the multimedia effect on affective-motivational variables. However, affective-motivational variables can also be investigated as moderators. A stronger effect of RPs on performance was found for students who reported a higher overall test-taking pleasure (Lindner et al., 2018) and students who expressed a higher appreciation for the domain of math (Cooper et al., 2018). It is conceivable that a high level of engagement allows test-takers to exploit the potential cognitive facilitation of multimedia material (especially RPs) more effectively. Yet, to make causal interpretations, studies should aim to experimentally vary the motivation of students to take a test with multimedia manipulations.

Gender In two studies, Dewolf et al. (2014) found no moderating effect of gender on the effects of DPs in testing. Consistent with this, gender was not a significant predictor of multimedia effects for RPs in a study (Lindner et al., 2018). This pattern is consistent with an assumption of general cognitive multimedia effects with no particular relation to gender. However, more research is needed in this area to obtain conclusive results.

Native Language Solano-Flores et al. (2014) investigated native English speakers and non-native speakers and found that the native language did not influence the effect of DPs. In contrast, Hartmann (2012) reported for items with IPs a stronger performance-enhancing effect for students who did not have the language of the test material as their first language. Studies that focus on the effects of RPs on non-native speakers would be desirable, as the dual coding of the picture information may mitigate language deficiencies in particular.

Moderators: Item-Level

Item-Complexity Understanding whether pictures have different effects depending on the item complexity is crucial for the targeted use of pictures in test items. The most elaborate study design so far was implemented by Wang et al. (2022). In a 3×3 within-subjects design, three types of pictures were combined with items of varying difficulty. The authors report that the item complexity interacted with the picture function. They conclude from their study, for example, that IPs may reduce accuracy on low-difficulty problems (due to unnecessary induced cognitive load) but could support students' understanding of more complex and difficult tasks. In line with this, other studies indicate that pictures can have a performance-enhancing effect especially for difficult items and this was found for IPs (Goolkasian, 1996; Hoogland et al., 2018b), RPs (Elia et al., 2007; Saß et al., 2012) and OPs (Múñez et al., 2013). However, Chu et al., (2017) found no considerable effect of varying complexities for OPs.

Subject Domain Lindner (2020) varied the subject domain and found the same pattern, i.e., multimedia effects for RPs but not for DPs, for both math and science items**.** However, studies with IPs suggest that effects of pictures may vary across different subdomains, such as in math (Hoogland et al., 2018b), medicine (Vorstenbosch et al., 2013), biology (Magnus et al., 2020) or programming (Whitley et al., 2006).

Dynamism and Modality Some studies have investigated whether dynamic pictures differ from static pictures regarding their multimedia effects. Two studies suggest a slight difference in item difficulty (Static > Dynamic; Wu et al., 2015 for IPs) and performance (Dynamic > Static; Ehrhart & Lindner, 2023 for RPs) for dynamic pictures compared to static pictures. Other studies found no particular benefits for dynamic as compared to static RPs (Dindar et al., 2015; Ehrhart et al., 2024). Moreover, the effects of RPs on problem-solving were not significantly influenced by text modality (e.g., Ehrhart & Lindner, 2023). As technology advances and computer-based testing becomes a standard, pictures with different dynamics, elements of interaction, and items with different text modality may become more commonly implemented (Arslan et al., 2020; Keehner et al., 2023). Thus, continuing research in this direction may be worthwhile.

Integration Format Several studies investigated how CTML instructional design principles (i.e., spatial contiguity: picture and text should be displayed in close proximity) affected performance in tasks and reported contradictory findings. Some studies (Dirkx et al., 2021 for IPs; Saß & Schütte, 2016 for IPs; Wang et al., 2022 for RPs) found that adjusting items according to common instructional design principles lowered item difficulty. One study found no moderating effect of the integration format for RPs and OPs (Saß et al., 2017). In contrast, Jarodzka et al. (2015) showed that students performed better on tasks with IPs in a split test item format compared to an integrated item format. However, they found no differences in visual transitions between fixations on text and picture for the two formats. This finding is still surprising and contradicts the spatial contiguity principle, which has received some empirical support also in the testing context (e.g., Moon et al., 2022). More research will be necessary to investigate if this effect is replicable across studies and if the effect may relate in a certain way to the nature of IPs.

Other Item-Design-Factors Carotenuto et al. (2021) found that changes in the presentation of a text-picture item, such as the way a question is posed, can affect students' suspension of sense making. Lindner et al. (2022) investigated the response format as a potential moderator for the effects of RPs and found a stronger effect in constructed-response items (i.e., open response format) compared with multiple-choice items (i.e., closed response format). Furthermore, Dewolf et al. (2017) found that highlighting contextual elements within DPs with an orange marker did not alter the occurrence of null effects.

Discussion

The present review had several objectives. The overall objective was to collect and organize the existing research on the use of multimedia in test contexts. A systematic overview was intended to help to better define the field, especially in distinction to multimedia learning. One specific goal of the review was to investigate whether different picture functions are associated with different effects. Similar attempts have also been made in multimedia learning (Carney & Levin, 2002; Levin et al., 1987) and multimedia testing (Hu et al., 2021; Lindner, 2021). However, the combination of a systematic literature search with a narrative presentation of the results is a new contribution to the field and allowed to consider the existing work in broader terms. Another contribution of this review is that it provides the first systematic overview of further potential moderator variables that have been studied in the context of multimedia testing (beyond the picture type). Lastly, our *narrative* approach allowed us to make another novel contribution to the field. So far, there has been no systematic investigation of the extent to which the proposed terminology for different picture functions is actually used. However, due to the importance of the picture type as a central moderator in multimedia testing (Hu et al., 2021; Lindner, 2020, 2021), it seems necessary to discuss and establish a uniform terminology in the field. Thus, we have placed particular emphasis on this aspect. Our proposal of an alternative and somewhat simplified terminology is presented in the discussion.

Multimedia Effects of Decorative Pictures The present review is the first that systematically considered the effects of decorative pictures on various outcome measures in multimedia testing. Our narrative synthesis indicates that decorative pictures do not affect performance, time on task, and metacognition to a meaningful extent. Although the chosen methodology was helpful for our purpose, it does limit the conclusions we can draw. A meta-analytic study of the effects of decorative pictures, which has not yet been done, would be a necessary complement. Further primary studies that can be included would be desirable. In particular, affective-motivational variables should be investigated, for which conclusions about multimedia effects can only be made carefully based on the identified studies so far.

Multimedia Effects of Representational and Organizational Pictures Our narrative review supports a previous meta-analysis (Hu et al., 2021) which found that representational pictures in test items have a significant effect on test-taker performance compared to text-only items. Our results indicate a slightly more stable effect for organizational pictures on performance compared to text-only. This suggests that, similar to multimedia learning (Schüler et al., 2019), the inclusion of pictures in multimedia tests can lead to higher accuracy rates particularly when spatial information is displayed and needs to be processed. However, as already discussed, it is difficult to distinguish between representational pictures and organizational pictures in existing studies in the field of multimedia testing, as organizational pictures could be understood as a specific type of representational pictures according to their definition.

We found heterogeneous data regarding the effects of representational pictures on time on task. For a deeper understanding of the cognitive processes involved, it would be important to conduct further eye-tracking studies. Although we identified seven eye-tracking studies already (Lindner et al., 2017a, b; Malone et al., 2020; Ögren et al., 2017; Ott et al., 2018; Saß et al., 2017; Wang et al., 2022; Zheng & Cook, 2012), some did not include a text-only control group (e.g., Jarodzka et al., 2015; Saß et al., 2017; Wang et al., 2022) which limits the informational value with regard to certain research questions. Moreover, a systematic review of eye-tracking studies in the field of multimedia testing, analogous to multimedia learning (Alemdag & Cagiltay, 2018), would be desirable in the future once the database is sufficient.

Our broad literature search and inclusion criteria allowed us to make solid assumptions about the effects of representational pictures on metacognitive and affective-motivational variables. In addition to enhanced performance, representational pictures seem to lead to both more confidence and more enjoyment in test taking. However, the database is still comparably small. In addition, future studies could use further established instruments to get a more nuanced overview of the effects of pictures on test-anxiety and other important achievement emotions (Pekrun et al., 2011). Future research may focus also on specific questions, such as the relationship between the certainty of responses and the actual performance and whether problems such as overconfidence can occur, as is also described in multimedia learning (Bjork et al., 2013). Moreover, in multimedia learning, there are attempts to clarify the relationship between performance and affective-motivational variables (e.g., Moreno & Mayer, 2007). Such efforts would also be important for future research in multimedia testing.

Multimedia Effects of Informational Pictures While we have drawn relatively straightforward conclusions regarding the effects of decorative, representational and organizational pictures on performance, this is not possible for informational pictures, given the present database. Remarkably, informational pictures constitute the only category for which significant performance-reducing effects were reported. The definition of informational pictures is comparatively broad and fosters this heterogeneity. For example, informational pictures can be used in the task stem (e.g., Hoogland et al., 2018a) as well as in the response options (e.g., Saß et al., 2012), which presumably places different demands on cognitive processing and could differentially influence the use of textual information as a guide for mental model construction (Schüler et al., 2019). Moreover, negative effects of informational pictures can be observed mainly in mathematics tasks solved by very young students (e.g., Berends & van Lieshout, 2009; Elia et al., 2007; Yang & Huang, 2004). Future research should ask to what extent informational pictures are appropriate for younger students and what role representational competence (Huinker, 2015) plays, for example, in different age groups and across cultures (Mayer et al., 1991).

Moreover, it should be noted that the research on informational pictures inevitably differs from the research on the other categories of pictures in the following aspect: With decorative, representational and organizational pictures it is easy to add a picture to a text-only condition. This creates two typical comparison conditions that differ only in the presence of the picture. In most studies, such a comparison is made between a text-only and a text-picture condition. This is different for informational pictures; by definition, they provide information that goes beyond the information provided in the text. Thus, for informational pictures there is often a comparison made between a text-only and a picture-only condition (e.g., Hoogland et al., 2018a) or the text in the text-only condition differs from the text in the text-picture condition (Berends & van Lieshout, 2009) because of the additional information in the informational pictures. Another distinctive aspect of informational pictures is that they provide information that is not already given in the text. An interesting question here would be whether the participants have been aware that they need to extract essential information from the pictures. It is possible that performance-reducing effects observed in some studies may have been due to the participants' lack of understanding this necessity. Whether an additional hint about the informative value of the picture has a moderating effect could be investigated experimentally in future studies. Moreover, further studies could investigate the cognitive processes involved in integrating text and picture information in test situations with informational pictures. Studies might discover which aspects prevent successful text-picture integration, such as time constraints during the test situation or a lack of awareness of the picture's informational content. Such findings could be a key to better understand the divergent results of studies with informational pictures.

Furthermore, there is a somewhat separate research tradition dealing with classical graphs (Pinker, 1990), such as bar graphs or pie charts which are frequently included into test items and problem-solving tasks (Strobel et al., 2018) and may often be classified as informational pictures. Future work may also reflect more on the relation of classical graphs and informational pictures to further expand our understanding of the effects of visual representations in test items.

Moderators Beyond the Picture Function Another contribution of this review is that it provides a first systematic overview of potential moderator variables that have been studied in the context of multimedia testing. This proved to be very difficult due to the sparse empirical data available. Moreover, it should be noted that the systematic identification of potential moderator variables is not without difficulties. While variables that have been explicitly varied are easy to identify in articles (e.g., Lindner et al., 2022; Solano-Flores et al., 2014), this is not always the case when variables have been analyzed as covariates or predictors (e.g., Cooper et al., 2018). Nonetheless, we would like to make some further suggestions on which moderator variables future studies could focus on.

Some studies identified for our review indicate that picture effects can be especially expected for more complex items. However, more research with larger sample sizes is needed. Future studies should more explicitly and transparently vary the complexity of items. Ideally, the variation in item difficulty is combined with a variation in picture function, as a stronger effect was only observed for representational and informational but not for decorative pictures in one study (Wang et al., 2022). Recent research should be considered to adequately define and measure item and task complexity, for example, building on the concept of element interactivity (e.g., Chen et al., 2023).

The integration format of text and pictures in test items is another interesting variable for practical test design. It can be related to the spatial contiguity principle of multimedia learning, which suggests that learners understand and retain information more effectively when related text and pictures are presented in close proximity. For multimedia testing, however, the body of research regarding this factor is very thin. It would be especially interesting to investigate informational pictures in this regard. Informational pictures provide relevant information that is complementary to the information in the text. It is possible that a complementary picture could be better perceived and processed by test takers when a text and the picture are displayed in a spatially non-integrated format, which was found in one particular study (Jarodzka et al., 2015). However, experimental studies that systematically vary both the picture function and the integration format are needed to replicate and further investigate this unexpected relation and consider alternative effects pattern.

One of the most interesting moderating factors at the student level is test takers' prior knowledge and general cognitive abilities, which may also be connected to student age. Only few primary studies included such analyses, and they also report somewhat conflicting results. Heterogeneity in the operationalization of student skills further adds to this complexity. Some studies have focused on domain-specific skills (e.g., Berends & van Lieshout, 2009; Cooper et al., 2018), others used a more general measure, such as the school track attended (e.g., Ehrhart & Lindner, 2023; Lindner et al., 2022). Overall, specific attention should be paid in the future to investigate test-taker characteristics in multimedia testing as potential moderating factors.

Proposal of a Hierarchical Picture Taxonomy for Multimedia Testing

Based on the included studies, we could show that the proposed terminology for picture classification (Elia & Philippou, 2004; Elia et al., 2007; Hu et al., 2021; Lindner, 2021) is not used very consistently in the literature. In particular, the term "informational" is rarely used, which may be due to the fact that the term does not exist in multimedia learning. The infrequent use of the term "informational" in primary studies may also be due to its lack of intuitiveness. To describe the picture function more intuitively, we would like to propose using the term **informative picture** instead. This term is more in line with the characteristic role of providing essential information and informing the test taker.

Following Elia and Philippou (2004) and their initial definition of organizational pictures (i.e., that they provide clues that support the solution procedure), it is difficult to distinguish organizational from representational pictures in the existing primary studies. An extension of the definition was suggested by Lindner (2021), namely to refer to organizational when the picture focuses on providing (abstract or schematic) visuo-spatial information that supports the solution process. However, the difference between organizational and representational pictures in existing studies remains highly related to the nature of the task. Even more, organizational pictures also fulfil the functional conditions for representational pictures, namely that they double-code information from the text. Therefore, we propose that organizational pictures should be considered a specific subcategory of representational pictures in future research, especially considering that organizational pictures have rarely been studied and have hardly been referred to as such in primary studies (see Table 1).

Taken together as displayed in Fig. 6, we propose that three categories (decorative, representational and informative), which are fundamentally different from each other, could be sufficient for a functional picture description in multimedia testing which would simplify further communication in the field. Accordingly, pictures that do not provide the reader with any information relevant to the solution of the task, should be classified as *decorative pictures* (DPs). Contextual decorative pictures may still have a relation to the content of the item text, but only in the sense that they visualize an element or scene from the general context. In contrast, pictures that depict solution-relevant information from the task text should be referred to as *representational pictures* (RPs). As an important distinguishing aspect, tasks with a representational picture can be solved even if a test taker does not process the picture, because all important information can be found in the text as well. This does not apply to informative pictures (IPs) that can be identified by evaluating if the solution of an item is possible without taking the picture information into account. If not, it is a clear indicator for an informative picture that represents solution-relevant, complementary information to the text.

Still, focusing solely on the functional aspect of a picture can dismiss other distinguishing characteristics. Thus, we would like to suggest an extended, stepwise, hierarchical taxonomy to allow for a better classification of potential moderating factors of pictures in multimedia testing (see Fig. 6). As a first layer, the picture function should be coded as described above (decorative, representational or informative). The functional differentiation of images is best suited as a first layer because—unlike other characteristics—it provides clearly separable and disjoint categories. However, subcategories may be useful to further differentiate picture functions. As mentioned earlier, representational pictures that have a primary focus on spatial information could be described as organizational. Decorative pictures could be distinguished into pictures that match the item context and pictures that are completely unrelated to the task context.

As a second layer, multiple other picture characteristics may be coded. Aspects to describe a picture could be, for example, the level of picture abstractness (e.g., iconic, schematic), the picture complexity (e.g., low, moderate, high), the picture type (e.g., photorealistic, comic), the colorfulness (e.g., colorful, black-white, grey), the dynamism (e.g., static, animated) or the level of interaction (e.g., interactive, non-interactive).

Our hierarchical approach has the potential to foster a more nuanced consideration of multimedia effects in testing in the future. In the context of our

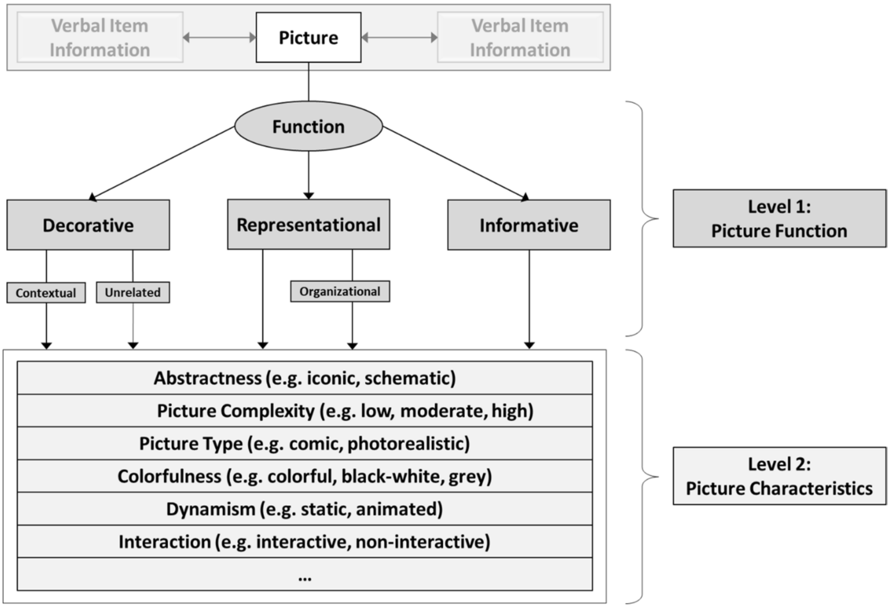

Verbal Item Information
Picture
Verbal Item Information
Function
Decorative
Representational
Informative
Contextual
Unrelated
Organizational
Level 1: Picture Function
Abstractness (e.g. iconic, schematic)
Picture Complexity (e.g. low, moderate, high)
Picture Type (e.g. comic, photorealistic)
Colorfulness (e.g. colorful, black-white, grey)
Dynamism (e.g. static, animated)
Interaction (e.g. interactive, non-interactive)
...
Level 2: Picture Characteristics

Fig. 6 Proposal of a Hierarchical Taxonomy for a Descriptive Classification of Picture Types in Multimedia Testing. **Note.** In our present work, we have so far referred to the term *informational* pictures as compared to *informative* pictures in this taxonomy model. The proposal of renaming this picture type is part of our endeavour to enhance and simplify the taxonomy for picture classification in multimedia testing

current review, however, it would not add information depth, as the primary studies and theoretical papers have not reported such additional aspects in most cases. Nevertheless, the second layer categories will require further careful definitions that go beyond the scope of this review. For example, at what point is an image considered abstract or image complexity classified as low, medium, or high? Addressing this type of open questions and extending the report of picture characteristics (and participant characteristics) in future primary studies would be an important aspect to further advance the field.

Limitations and Conclusion

This review is subject to some limitations that are important to note. We deliberately chose very broad inclusion criteria to evaluate as many potentially relevant studies as possible from the still young and small research field. Accordingly, the methodological qualities and the design of the individual studies are very heterogeneous and not in the focus of the present work. Therefore, an appropriate calculation of effect sizes that would be comparable across research studies is not possible without losing an extensive number of studies. We are aware that relying on vote counting can lead to misinterpretation of the data (Borenstein et al., 2021). However, for this first systematic narrative review, which was intended to provide a comprehensive overview of the entire field, it seemed to be a valuable approach. We are currently working on a complementary meta-analysis to provide quantitative support for the findings of this systematic narrative review.

Nevertheless, the present review is an important contribution to the research on multimedia testing. On the one hand, it sets important impulses for the delineation and standardization of the research field. On the other hand, it may support educators and test designers contemplating the use of pictures in their materials.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by the Leibniz Association, Germany (Leibniz Competition 2019), for the Leibniz Junior Research Group COMET (Cognition and Motivation in Educational Testing; Grant Number J55/2018, awarded to Marlit Lindner, PI).

Data Availability The data are available upon request from the first author.

Declarations

Conflict of Interest The authors declare that they have no financial interests or personal relationships that could have influenced the work reported in this paper.

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.), R*esearch in mathematics education: Conference of Five Cities: Nicosia, Rhodes, Bologna, Palermo, Locarno,* 39–59. Cyprus: University of Cyprus.

*Ahmed, A., Hurwitz, D., Gestson, S., & Brown, S. (2021). Differences between professionals and students in their visual attention on multiple representation types while solving an open-ended engineering design problem. *Journal of Civil Engineering Education, 147*(3), 04021005 https://doi.org/10.1061/(ASCE)EI.2643-9115.0000044

Ainsworth, S. (1999). The Functions of Multiple Representations. *Computers & Education, 33*(2–3), 131–152. https://doi.org/10.1016/S0360-1315(99)00029-9

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

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education, 125*, 413–428. https://doi.org/10.1016/j.compedu.2018.06.023

Alesandrini, K. L. (1984). Pictures and Adult Learning. *Instructional Science, 13*(1), 63–77. https://doi.org/10.1007/BF00051841

Arslan, B., Jiang, Y., Keehner, M., Gong, T., Katz, I. R., & Yan, F. (2020). The effect of drag-and-drop item features on test-taker performance and response strategies. *Educational Measurement: Issues and Practice, 39*(2), 96–106. https://doi.org/10.1111/emip.12326

Atkinson, C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American, 225*(2), 82–91.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556–559. https://doi.org/10.1126/science.1736359

*Bahlmann, O. (2018). Illustrated versus non-illustrated anatomical test items in anatomy course tests and german medical licensing examinations (M1). *GMS Journal for Medical Education, 35*(2), Doc25 (20180515). https://doi.org/10.3205/ZMA001172

*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*, 345–353. https://doi.org/10.1016/j.learninstruc.2008.06.012

*Beveridge, M., & Parkins, E. (1987). Visual representation in analogical problem solving. *Memory & Cognition, 15*(3), 230–237. https://doi.org/10.3758/BF03197721

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

*Booth, J. L., & Koedinger, K. R. (2012). Are diagrams always helpful tools? Developmental and individual differences in the effect of presentation format on student problem solving. *British Journal of Educational Psychology, 82*, 492–511. https://doi.org/10.1111/j.2044-8279.2011.02041.x

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2021). *Introduction to Meta-Analysis* (2nd ed.). John Wiley & Sons.

*Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology, 23*(3), 369–381. https://doi.org/10.1002/acp.1460

Carney, R. N., & Levin, J. R. (2002). Pictorial Illustrations Still Improve Students' Learning from Text. *Educational Psychology Review, 14*(1), 5–26. https://doi.org/10.1023/A:1013176309260

*Carotenuto, G., Di Martino, P., & Lemmi, M. (2021). Students' Suspension of Sense-Making in Problem Solving. *ZDM – Mathematics Education, 53*(4), 817–830. https://doi.org/10.1007/s11858-020-01215-0

Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction, 8*(4), 293–332. https://doi.org/10.1207/s1532690xci0804_2

Chen, O., Paas, F., & Sweller, J. (2023). A Cognitive Load Theory Approach to Defining and Measuring Task Complexity Through Element Interactivity. *Educational Psychology Review, 35*(2), 63. https://doi.org/10.1007/s10648-023-09782-w

*Chu, J., Rittle-Johnson, B., & Fyfe, E. R. (2017). Diagrams benefit symbolic problem-solving. British *Journal of Educational Psychology, 87*(2), 273–287. https://doi.org/10.1111/bjep.12149

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3*(3), 149–210. https://doi.org/10.1007/BF01320076

*Clinton, V., & Walkington, C. (2019). Interest-enhancing approaches to mathematics curriculum design: illustrations and personalization. *The Journal of Educational Research, 112*(4), 495–511. https://doi.org/10.1080/00220671.2019.1568958

*Cooper, J. L., Sidney, P. G., & Alibali, M. W. (2018). Who benefits from diagrams and illustrations in math problems? ability and attitudes matter. *Applied Cognitive Psychology, 32*(1), 24–38. https://doi.org/10.1002/acp.3371

*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

*Dewolf, T., Van Dooren, W., Cimen, E. E., & Verschaffel, L. (2014). The Impact of Illustrations and Warnings on Solving Mathematical Word Problems Realistically. *The 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 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., Kabakçı Yurdakul, I., & Dönmez, F. I. (2015). Measuring cognitive load in test items: static graphics versus animated graphics. *Journal of Computer Assisted Learning, 31*(2), 148–161. https://doi.org/10.1111/jcal.12086

*Dirkx, K. J. H., Skuballa, I., Manastirean-Zijlstra, C. S., & Jarodzka, H. (2021). Designing computer-based tests: design guidelines from multimedia learning studied with eye tracking. *Instructional Science, 49*(5), 589–605. https://doi.org/10.1007/s11251-021-09542-9

*Ehrhart, T., & Lindner, M. A. (2023). Computer-Based Multimedia Testing: Effects of Static and Animated Representational Pictures and Text Modality. *Contemporary Educational Psychology, 73*. Article 102151. https://doi.org/10.1016/j.cedpsych.2023.102151

*Ehrhart, T., Höffler, T., Grund, S., & Lindner, M. A. (2024). Static versus dynamic representational and decorative pictures in mathematical word problems: Less might be more. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000821

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, Norway: 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

*Fırat, M. (2017). How real and model visuals affect the test performance of elementary students. *Computers in Human Behavior, 71*, 258–265. https://doi.org/10.1016/j.chb.2017.02.021

*Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine, 83*, 27–33. https://doi.org/10.1016/j.socscimed.2013.01.034

*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

*Garret, A. J. (2008). *The role of picture perception in children's performance on a picture vocabulary test* [Doctoral Dissertation, University of Maryland]. ProQuest Dissertations & Theses Global. Retrieved December 10, 2022, from https://www.proquest.com/docview/304425024/

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

*Goolkasian, P. (2000). Pictures, words, and sounds: from which format are we best able to reason? *The Journal of General Psychology, 127*(4), 439–459. https://doi.org/10.1080/00221300009598596

*Gray, K., Owens, K., Liang, X., & Steer, D. (2012). Assessing multimedia influences on student responses using a personal response system. *Journal of Science Education and Technology, 21*(3), 392–402. https://doi.org/10.1007/s10956-011-9332-1

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ, 331*(7524), 1064–1065. https://doi.org/10.1136/bmj.38636.593461.68

*Günbaş, N. (2020). Students solve mathematics word problems in animated cartoons. *Malaysian Online Journal of Educational Technology, 8*(2), 43–57.

*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

*Hartmann, S. (2012). *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 the achievement of written tests for the assessment of scientific competence). (Doctoral Dissertation, Universität Duisburg-Essen). Retrieved December 16, 2022, from https://duepublico2.uni-due.de/servlets/MCRFileNodeServlet/duepublico_derivate_00033260/hartmann_diss.pdf

Hegarty, M., Carpenter, P. A., & Just, M. A. (1991). Diagrams in the Comprehension of Scientific Texts. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research, Vol. 2*. (pp. 641–668). Lawrence Erlbaum Associates, Inc.

*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*, 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

Hu, L., Chen, G., Li, P., & Huang, J. (2021). Multimedia effect in problem solving: a meta-analysis. *Educational Psychology Review, 33*(4), 1717–1747. https://doi.org/10.1007/s10648-021-09610-z

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

*Hunt, D. R. (1978). Illustrated multiple choice examinations. Medical Education, 12, 417–420.

*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

Keehner, M., Arslan, B., & Lindner, M. A. (2023). Cognition-centered design principles for digital assessment tasks and items. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International Encyclopedia of Education* (4th ed., pp. 171–184). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10025-9

Kirschner, P. A., Park, B., Malone, S., & Jarodzka, H. (2017). Toward a cognitive theory of multimedia assessment (CTMMA). In M. J. Spector, B. B. Lockee, & M. D. 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. https://doi.org/10.1016/j.cedpsych.2006.07.002

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

Levin, J. R., Anglin, G. J., & Carney, R. N. (1987). On Empirically Validating Functions of Pictures in Prose. *The Psychology of Illustration, 1*, 51–86.

Levin, J. R. (1981). On functions of pictures in Prose. In F. J. Pirozzolo, & M. C. Wittrock (Eds.), *Neuropsychological and Cognitive Processes in Reading* (pp. 203–228). New York: Academic Press. https://doi.org/10.1016/B978-0-12-185030-2.50013-5

*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., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: a matter of motivation and cognitive resources. *Frontiers in Psychology, 10*, 1533. https://doi.org/10.3389/fpsyg.2019.01533

*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., Ihme, J. M., Saß, S., & Köller, O. (2018). 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., Barenthien, J., & Köller, O. (2021). 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

*Lindner, M. A., Schult, J., & Mayer, R. E. (2022). A Multimedia effect for multiple-choice and constructed-response test items. *Journal of Educational Psychology, 114*(1), 72–88. https://doi.org/10.1037/edu0000646

*Lindner, M. A. (2020). Representational and decorative pictures in science and mathematics tests: do they make a difference? *Learning and Instruction, 68*, 101345. https://doi.org/10.1016/j.learninstruc.2020.101345

Lindner, M. A. (2021). Principles for Educational Assessment with Multimedia. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge Handbook of Multimedia Learning* (3rd ed., pp. 552–565). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108894333.055

*Magnus, L., Schütte, K., & Schwanewedel, J. (2020). Challenges solving science tasks with text–picture combinations persist beyond secondary school. *Journal of Research on Educational Effectiveness, 13*(4), 759–783.https://doi.org/10.1080/19345747.2020.1750744

*Malone, S., Altmeyer, K., Vogel, M., & Brünken, R. (2020). Homogeneous and heterogeneous multiple representations in equation-solving problems: an eye-tracking studyhomogeneous and heterogeneous multiple representations in equation-solving problems: an eye-tracking study. *Journal of Computer Assisted Learning, 36*(6), 781–798. https://doi.org/10.1111/jcal.12426

Mayer, R. E., Tajika, H., & Stanley, C. (1991). Mathematical Problem Solving in Japan and the United States: A Controlled Comparison. *Journal of Educational Psychology, 83*(1), 69–72. https://doi.org/10.1037/0022-0663.83.1.69

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (1st ed., pp. 31–48). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.004

Moon, J. A., Lindner, M. A., Arslan, B., & Keehner, M. (2022). Investigating the Split-Attention Effect in Computer-Based Assessment: Spatial Integration and Interactive Signaling Approaches. *Educational Measurement: Issues and Practice, 41*(2), 90–117. https://doi.org/10.1111/emip.12485

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

*Múñez, D., Orrantia, J., & Rosales, J. (2013). The effect of external representations on compare word problems: supporting mental model construction. *The Journal of Experimental Education, 81*(3), 337–355. https://doi.org/10.1080/00220973.2012.715095

Noetel, M., Griffith, S., Delaney, O., Harris, N. R., Sanders, T., Parker, P., del Pozo Cruz, B., & Lonsdale, C. (2022). Multimedia Design for Learning: An Overview of Reviews with Meta-Meta-Analysis. *Review of Educational Research, 92*(3), 413–454. https://doi.org/10.3102/00346543211052329

Novick, L. R., & Bassok, M. (2005). Problem Solving. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 321–349). Cambridge University Press.

*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

Organization for Economic Co-operation and Development (OECD). (2009). *PISA 2006 Technical Report*. OECD Publishing. https://doi.org/10.1787/9789264048096-en

*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*, 88–105. https://doi.org/10.1016/j.learninstruc.2018.04.010

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. Oxford University Press.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: the achievement emotions questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002

Pinker, S. (1990). A Theory of Graph Comprehension. In R. O. Freedle (Ed.), *Artificial Intelligence and the Future of Testing* (pp. 73–126). Lawrence Erlbaum Associates.

*Ramjan, L. M. (2011). Contextualism adds realism: nursing students' perceptions of and performance in numeracy skills tests. *Nurse Education Today, 31*(8), e16–e21. https://doi.org/10.1016/j.nedt.2010.11.006

*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: pictures and item processing. *Applied Cognitive Psychology, 26*(1), 70–81. https://doi.org/10.1002/acp.1798

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

Scheiter, K., & Eitel, A. (2015). Signals foster multimedia learning by supporting integration of highlighted text and diagram elements. *Learning and Instruction, 36*, 11–26. https://doi.org/10.1016/j.learninstruc.2014.11.002

Scheiter, K., Schüler, A., Gerjets, P., Huk, T., & Hesse, F. W. (2014). Extending multimedia research: how do prerequisite knowledge and reading comprehension affect learning from text and pictures? *Computers in Human Behavior, 31*, 73–84. https://doi.org/10.1016/j.chb.2013.09.022

*Schnotz, W., & Wagner, I. (2018). Construction and elaboration of mental models through strategic conjoint processing of text and pictures. *Journal of Educational Psychology, 110*(6), 850–863. https://doi.org/10.1037/edu0000246

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

*Schnotz, W., Ludewig, U., Ullrich, M., Horz, H., McElvany, N., & Baumert, J. (2014). Strategy shifts during learning from texts and pictures. *Journal of Educational Psychology, 106*(4), 974–989. https://doi.org/10.1037/a0037054

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. https://doi.org/10.1111/bjop.12341

Slough, S. W., & McTigue, E. (2013). Development of the graphical analysis protocol (GAP) for eliciting the graphical demands of science textbooks. In M. S. Khine (Ed.), *Critical Analysis of Science Textbooks* (pp. 17–30). Springer. https://doi.org/10.1007/978-94-007-4168-3_2

*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.https://doi.org/10.1080/10627197.2014.964116

*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. https://doi.org/10.1080/15305058.2015.1099534

Strobel, B., Lindner, M. A., Saß, S., & Köller, O. (2018). Task-irrelevant data impair processing of graph reading tasks: an eye tracking study. *Learning and Instruction, 55*, 139–147. https://doi.org/10.1016/j.learninstruc.2017.10.003

*Strobel, B., Grund, S., & Lindner, M. A. (2019). Do seductive details do their damage in the context of graph comprehension? Insights from eye movements. *Applied Cognitive Psychology, 33*(1), 95–108. https://doi.org/10.1002/acp.3491

*Vorstenbosch, M. A. T. M., Klaassen, T. P. F. M., Kooloos, J. G. M., Bolhuis, S. M., & Laan, R. F. J. M. (2013). Do images influence assessment in anatomy? Exploring the effect of images on item difficulty and item discrimination. *Anatomical Sciences Education, 6*(1), 29–41. https://doi.org/10.1002/ase.1290

*Wang, X., Kang, W., Huang, L., & Li, L. (2022). The impact of illustrations on solving mathematical word problems for chinese primary school students: evidence for a split attention effect on eye-movement research. *ZDM – Mathematics Education, 54*(3), 555–567. https://doi.org/10.1007/s11858-022-01357-3

*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

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

*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 & Education, 85*, 35–48. https://doi.org/10.1016/j.compedu.2015.01.007

*Yang, D., & Huang, F. (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

*Zheng, R., & Cook, A. (2012). Solving complex problems: a convergent approach to cognitive load measurement. *British Journal of Educational Technology, 43*(2), 233–246. https://doi.org/10.1111/j.1467-8535.2010.01169.x

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

Authors and Affiliations

Lauritz Schewior[1] · **Marlit Annalena Lindner**[1,2]

✉ Lauritz Schewior
schewior@leibniz-ipn.de

Marlit Annalena Lindner
mlindner@leibniz-ipn.de

[1] IPN - Leibniz Institute for Science and Mathematics Education, Kiel, Germany

[2] IWM - Leibniz Institut für Wissensmedien, University of Tübingen, Tübingen, Germany