Journal of Educational Psychology

© 2024 The Author(s)
ISSN: 0022-0663

2024, Vol. 116, No. 4, 532–549
https://doi.org/10.1037/edu0000821

Static Versus Dynamic Representational and Decorative Pictures in Mathematical Word Problems: Less Might Be More

Tom Ehrhart[1], Tim Niclas Höffler[1], Simon Grund[2], and Marlit Annalena Lindner[1, 3]
[1] IPN—Leibniz Institute for Science and Mathematics Education, University of Kiel
[2] Institute of Psychology, University of Hamburg
[3] IWM—Leibniz Institut für Wissensmedien, University of Tübingen

Research on the multimedia effect in testing indicates that static representational pictures (RPs) and, potentially, dynamic RPs that further subdivide the picture into segments may support students' mental processing. This might be especially relevant for mathematical word problems that pose high mental demands in a multistage solution process. Existing studies further indicate that contrary to expectations of practitioners, static decorative pictures (DPs) do not improve students' affective state. It is unclear if dynamic DPs that decorate each segment of a word problem better meet these expectations. In our preregistered online experiment that involved 308 students in a 3 × 2 mixed design, we manipulated word problems regarding three visualization conditions (RPs vs. DPs vs. text-only), and two kinds of dynamics (static vs. dynamic). As expected, static and dynamic RPs increased response correctness, metacognitive ratings, and satisfaction compared to text-only. Besides this replication and extension of the multimedia effect in testing, dynamic RPs did not outperform static RPs in direct comparisons, however. Both RP conditions did not extend the time-on-task compared to text-only. As expected, static DPs did neither increase response correctness nor metacognition nor satisfaction compared to text-only. However, dynamic DPs were also unable to increase satisfaction compared to text only. Additional analyses that took the item position into account unveiled that the time-on-task in dynamic DP items aligned to that of text-only items over the course of the experiment, so that the students might have ignored dynamic DPs over time. Finally, we conclude implications for using these visualizations in digital assessments.

Educational Impact and Implications Statement

Digital assessments provide opportunities to incorporate pictures and complex visualizations into test items. However, is it worthwhile to invest in creating dynamic visualizations instead of static ones? This experiment demonstrates that both static and dynamic representational pictures (RPs) have a positive impact on student performance, metacognition, and satisfaction. Yet, the dynamic design of pictures did not offer additional benefits compared to static RPs. Additionally, neither static nor dynamic decorative pictures enhanced students' performance, metacognition, or satisfaction during the test. Overall, our findings suggest that practitioners should carefully evaluate the potential benefits and costs of including dynamic and decorative multimedia elements in educational tasks, as the additional construction efforts may not yield significant returns.

Keywords: multimedia effect in testing, representational pictures, decorative pictures, animations, word problems

Supplemental materials: https://doi.org/10.1037/edu0000821.supp

This article was published Online First February 26, 2024.
Olusola O. Adesope served as action editor.
Tom Ehrhart https://orcid.org/0000-0002-7735-9519
The study design and hypotheses were preregistered (see https://aspredicted.org/GF9_T5J). Data are available upon request from Tom Ehrhart. Additional materials are provided in an online repository (see https://osf.io/z746n/?view_only=a28044d443ce443bb496ed97a7ac0438). 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 J55/2018, awarded to Marlit Annalena Lindner, principal investigator).

Tom Ehrhart served as lead for data curation, formal analysis, investigation, software, visualization, and writing–original draft. Tim Niclas Höffler served in a supporting role for project administration and supervision. Simon Grund served in a supporting role for formal analysis, software, and writing–review and editing. Marlit Annalena Lindner served as lead for funding acquisition, resources, supervision, and writing–review and editing and served in a supporting role for investigation and software. Tom Ehrhart and Tim Niclas Höffler contributed equally to writing–review and editing. Tom Ehrhart and Marlit Annalena Lindner contributed equally to methodology and project administration. Tom Ehrhart, Tim Niclas Höffler, Simon Grund, and Marlit Annalena Lindner contributed equally to validation.

The preregistered design is available at https://aspredicted.org/GF9_T5J.
Correspondence concerning this article should be addressed to Tom Ehrhart, IPN—Leibniz Institute for Science and Mathematics Education, University of Kiel, Olshausenstraße 62, 24118 Kiel, Germany. Email: ehrhart@leibniz-ipn.de

In educational test items, different picture types are used to adapt the task characteristics; this is based on the idea that a visual representation besides the text can help test takers to process the information concurrently in two cognitive channels (Mayer, 2014; Schnotz & Bannert, 2003). Building on such cognitive facilitating effects, pictures can help to adapt item difficulty to create items for the entire proficiency spectrum. With regard to affective-motivational effects, it is desirable to design test items that increase test-taking satisfaction to counteract issues with test-taking motivation, which often threatens a valid interpretation of test scores in low-stakes assessments (e.g., see Trendtel & Robitzsch, 2018). In mathematical word problems—a very popular item type—visualizations can be used to manage the cognitive demands during the mental modeling process (e.g., see Blum & Leiß, 2007; Verschaffel et al., 2020).

Among the different picture types, representational pictures (RPs) are most commonly used. RPs visualize essential content from the text with the intention to facilitate the task understanding (Carney & Levin, 2002). Therefore, RPs and text act as multiple representations. As an adaption of the multimedia effect in learning, research on the multimedia effect in testing has shown positive effects of RPs on students' cognitive and affective-motivational item processing (see, e.g., Hu et al., 2021; Lindner, 2021, for reviews). These beneficial effects of RPs were not only found for static RPs that visualize all representational elements at once (e.g., Lindner, 2020) but also for segmented RPs that fade in the pictorial elements stepwise over time (i.e., dynamic RPs, see Figure 1). Yet, beneficial effects of dynamic RPs compared to static RPs and text-only items have only been shown in one single experiment so far (Ehrhart & Lindner, 2023), and should be further investigated.

In contrast to RPs, decorative pictures (DPs) do not visualize essential task information but instead represent the broader topic of the task in an appealing way. DPs are often employed in testing with the intention to increase students' test-taking enjoyment (Lindner, 2020), but research on the effects of DPs in testing is still rare (Lindner, 2021). The small number of experimental studies conducted focused exclusively on static DPs that provide a single DP for an entire word problem (see Hu et al., 2021); for such static DPs, there is no indication of affective-motivational benefits in test items so far (e.g., Berends & van Lieshout, 2009; Lindner, 2020). Yet, as static DPs do not decorate the entire storyline of a word problem, it might be more appealing to students if several DPs are provided in a slideshow to decorate different content segments of a word problem (i.e., dynamic DPs; see Figure 2). Effects of such dynamic DPs have not been researched so far.

Against this background, the present study targeted two main areas. First, it directly connects to previous research on static visualizations in testing as it aimed to replicate the beneficial effects of static RPs in word problems and to provide further evidence with regard to the apparent absence of affective-motivational effects from static DPs (Lindner, 2020). Second, it extends knowledge about segmented dynamic visualizations in tests, which can be considered as a link between conventional static pictures and animations. This picture type is inspired by a common heuristic of students to mentally segment word problems (Verschaffel et al., 2020). A specific aim of the present study was therefore to evaluate whether and, if so, to what extent dynamic RPs and dynamic DPs may be more beneficial in terms of cognitive, metacognitive, and affective-motivational effects as compared to their static counterparts. Knowledge on the effects of different picture types can help to derive specific design recommendations for multimedia test items in the long run (see Kirschner et al., 2017).

Mental Model Construction in Multimedia Tasks

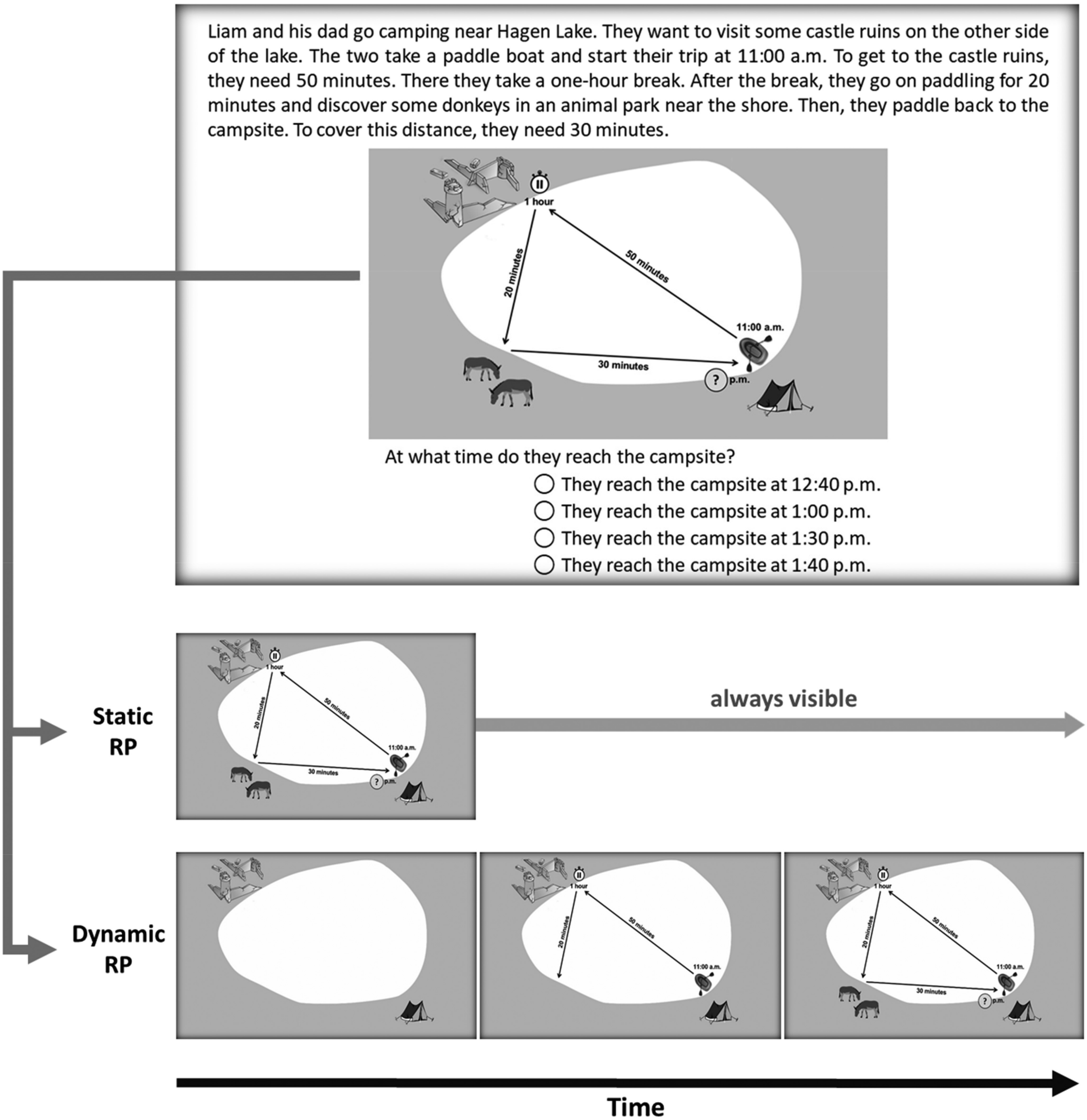
When students work on educational tasks, they need to actively deal with the presented content to build a mental model and apply their existing knowledge to the task (Mayer, 2014; Schnotz, 2005). The dual coding theory (Paivio, 1986) assumes that two different channels of limited capacity can be used to process the content at the same time, namely, the visual and the auditory channels. Building on this, the cognitive theory of multimedia learning (CTML; Mayer, 2014) assumes that a combination of text and visualizations facilitates the cognitive processing in educational tasks compared to text alone (i.e., multimedia effect). In the CTML, the channels are described as visual-pictorial and auditory-verbal, which organize the information in two separate mental models before integrated into one common model. Likewise, the integrated model of text and picture comprehension (ITPC; Schnotz & Bannert, 2003) distinguishes between depictive content (i.e., visualizations; expected to be directly understandable), and descriptive content (i.e., written or spoken text; expected to be encrypted so that it needs to be transformed into a depictive format before it can be integrated). Building on both the CTML and the ITPC, it can be assumed that additional visualizations in a task can reduce the cognitive demands placed on students when they are solving an item (see the cognitive load theory, CLT; Chandler & Sweller, 1991) given that the added pictures are designed to enhance the task presentation (e.g., double-coding of task-relevant information in RPs). However, pictures that do not convey task-relevant information (e.g., DPs) are not expected to have similar positive effects on cognition and may even serve as seductive details (Harp & Mayer, 1997).

The cognitive-affective theory of learning with media (CATLM; Moreno & Mayer, 2007) further extends the CTML by also hypothesizing an affective-motivational level. It elaborates that metacognitive improvements (e.g., from visualizations) can enhance students' affect and motivation during the mental modeling process, which can, consequently, improve the quality of the final mental model. The expectancy-value theory (Eccles & Wigfield, 2020) provides further explanations about how students' metacognitive ratings can impact affective-motivational outcomes. Accordingly, students estimate the task-effort costs to determine their expectancy of success (see Flake et al., 2015) and improve their affective-motivational attitude, if the task value is considered to be worth the costs. Thus, cognitive facilitations in test items that comprise text and visualizations might increase students' expectancy of success and also improve their item-solving satisfaction (e.g., Lindner, 2020; Lindner et al., 2018). Due to limited cognitive capacity, these potential enhancements from visualizations might be especially important for tasks that have high inherent complexity and require efficient mental processing of the content, such as, for example, mathematical word problems.

Mental Model Construction in Word Problems

Mathematical word problems can be defined as verbal descriptions of problem situations that pose mathematical questions and require the use of given data to perform calculations and answer the given questions (see Verschaffel et al., 2020). Due to the

Figure 1
Comparison of Static and Dynamic RPs in an Exemplary Word Problem

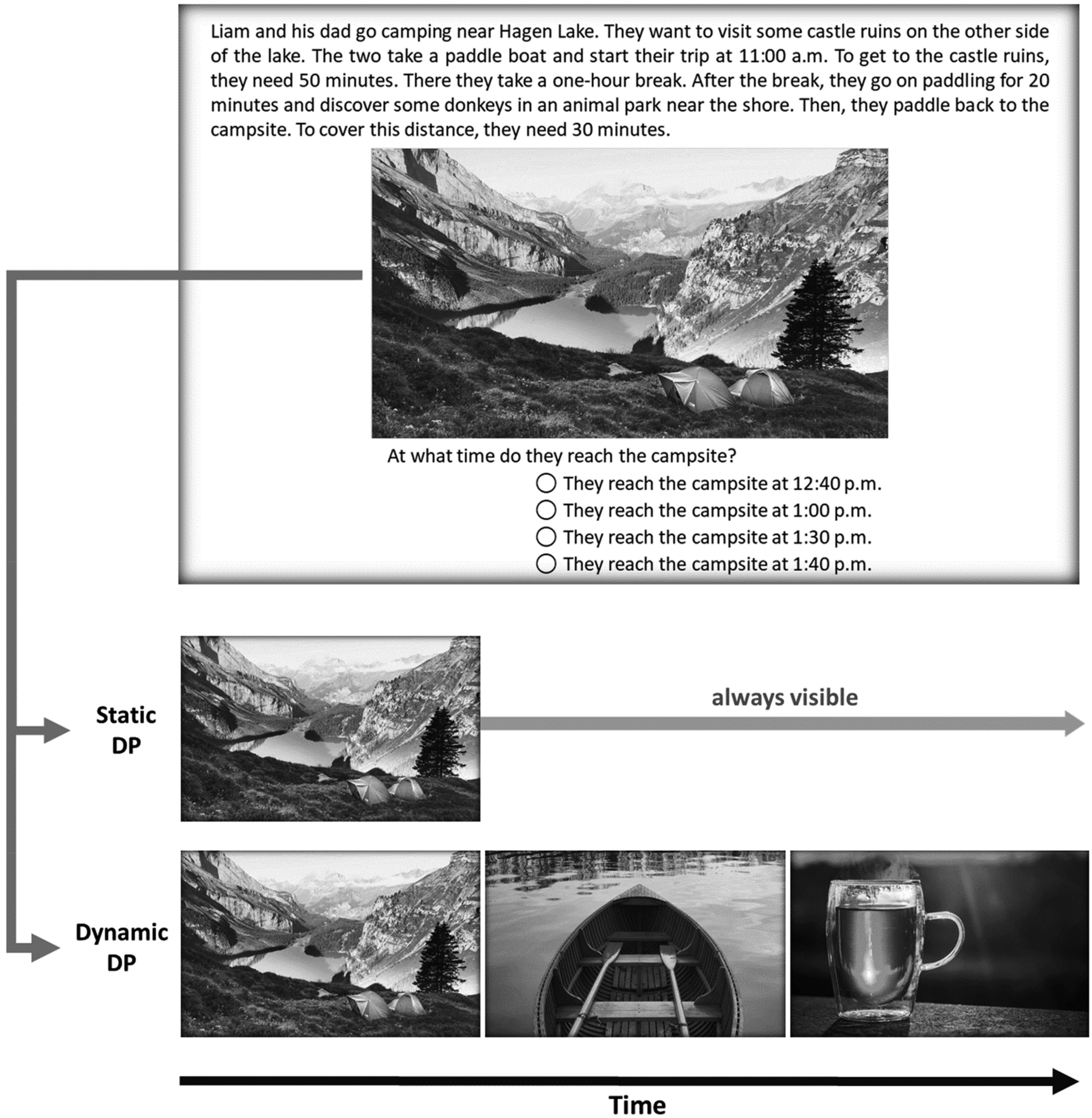


Note. Static RPs always visualized all pictorial elements. In dynamic RPs, the pictorial elements were faded in successively and complemented the visualization over time, so that they were finally identical to static RPs. RPs = representational pictures.

complex mental construction process they require, word problems can be seen as an important representative for test items with high cognitive demands. Solving these tasks requires students to build a mental model of the real-life context (i.e., situational model) before then transferring it to a mathematical context (i.e., mathematical model; Verschaffel et al., 2020).

Students' mental construction in these two contexts is further carried out in the modeling cycle of Blum and Leiss (2007): (a) Given the real situation, students initially build a mental representation (situational model). (b) Still within the situational context, the model gets simplified and structured by focusing on more essential details (real model). (c) Then the students transform these components into

Figure 2
Comparison of Static and Dynamic DPs in an Exemplary Word Problem



Note. Static DPs always visualized one static picture. In dynamic DPs, multiple static DPs were presented one after another in terms of a slideshow. DPs = decorative pictures.

a mathematical environment (mathematical model). (d) Within this environment, students can perform calculations and gain first results (mathematical results). (e) These results need to be transferred back to the situational context (real results). (f) The real results are then implemented into the situational model. In cases of low plausibility or when aspects remain unclear, the previous steps can be repeated until a satisfactory solution is found (validation). (g) Finally, the students refer their validated results to the real situation and decide on an answer (exposition). The modeling cycle indicates that the situational mental model underlies all subsequent models and that it has a high relevance at different stages of processing the task and solving the problem.

Accordingly, a wide range of publications underline the central role of the situational model in mathematical word problems (e.g., Jitendra et al., 2007; Leiss et al., 2019; Mayer & Hegarty, 1996; Verschaffel et al., 2020), providing evidence that it has a greater impact on performance than mathematical knowledge does (Fuchs et al., 2006; Stephany, 2021). The quality of the situational model depends on a variety of cognitive aspects, such as verbal and nonverbal problem-solving skills (e.g., Fuchs et al., 2006; Stephany, 2021), conceptual and procedural knowledge (Jitendra et al., 2007), and concept formation, which describes students' ability to identify, categorize, and determine rules in the task (Fuchs et al., 2006). On the basis of these skills, students can use many different strategies to build their situational models, such as building mental segmentations or producing graphical representations (Verschaffel et al., 2020).

RPs in Testing

In educational tests, different types of visualizations are used. RPs, as the most common kind of visualizations used in educational research (Carney & Levin, 2002), aim to support students' mental model construction by representing essential information from the item stem in a visual way. RPs do not add task-relevant information beyond the text information and act as multiple representations (see, e.g., Lindner, 2021), except for the additional information that may result from the natural characteristics of pictures (e.g., each object is presented in a more detailed and concrete manner compared to the text; relations between objects become explicit, which can reduce ambiguity as compared to the text information; see Ainsworth, 2006).

Empirically, several studies have found indications for the presence of the multimedia effect in testing regarding both students' performance and their metacognitive ratings, namely, their self-assessed expected performance. Moreover, in addition to the cognitive facilitations of RPs, several studies have also found positive effects of RPs on affective-motivational outcomes (i.e., item-solving satisfaction; Lindner, 2020; Lindner et al., 2018, 2022; Lindner, Lüdtke, et al., 2017).

Studies that provide evidence of a cognitive multimedia effect in testing have mainly been conducted in science (e.g., Lindner, 2020; Lindner et al., 2018; Saß et al., 2012) and mathematics (e.g., Cooper et al., 2018; Hoogland et al., 2018; Lindner, 2020; Orrantia et al., 2014; Ott et al., 2018). RPs in scientific and mathematical word problems seem to be especially important to support students in building appropriate mental models, which are required as a base to solve the problem. As shown by an eye-tracking study (Lindner, Eitel, et al., 2017), RPs in scientific test items were typically visited by students at an early stage of the task processing to gain an idea of the global structure of the task (i.e., as a mental scaffold) and were then revisited in the later course of the item processing during the problem-solving phase (i.e., mental model updating; Lindner, Eitel, et al., 2017). These inspections of RPs at the beginning and the end of the task processing seemed to reduce students' processing times of the accompanying text, as shown by the eye-tracking data.

With regard to the effects of RPs on time-on-task, however, the overall pattern of results is more mixed (see Hu et al., 2021). In science items, RPs typically did not change students' time-on-task (Lindner et al., 2022; Lindner, Eitel, et al., 2017; Lindner, Lüdtke, et al., 2017; Ott et al., 2018; Saß et al., 2012), whereas studies conducted in the context of mathematical word problems tentatively indicated that RPs can also increase time-on-task compared to text-only items (e.g., Lindner, 2020). This might be because of the fact that RPs in word problems need to represent both situational and mathematical elements (i.e., objects and numerals; Berends & van Lieshout, 2009; Lindner, 2020).

Connecting these findings to the context of mathematical word problems, early inspections of RPs, that is, at the beginning of the task, could improve the process of building an appropriate situational model. In the later course of the item processing, RPs could then support the integration of the real results into the situational model (see Blum & Leiß, 2007). Thus, RPs might have an important effect on cognitive processing in mathematical word problems (see, e.g., Cooper et al., 2018; Hoogland et al., 2018; Lindner, 2020). In contrast to conventional static RPs, a dynamic presentation could optimally segment the task content and might further facilitate students' mental processing of the task.

Dynamic RPs

Several meta-analyses have shown that representational visualizations that present the content in a dynamic way can improve the mental model construction as compared to static RPs (Berney & Bétrancourt, 2016; Höffler & Leutner, 2007; Ploetzner et al., 2020). These meta-analyses focused on the most common kind of dynamic visualizations, namely animations (Ploetzner & Lowe, 2012), defined as "any application which generates a series of frames, so that each frame appears as an alteration of the previous one, and where this sequence of frames is determined either by the designer or the user" (p. 782). Overall, the use of conventional animations has been especially recommended to demonstrate the functioning of dynamic systems or complex procedures with transitory changes (Berney & Bétrancourt, 2016). A reanalysis of these meta-analytic results by Ploetzner et al. (2020) indicated that conventional animations are especially effective to support the understanding of dynamic contexts; yet, conventional animations seem not optimal to support conceptual mental modeling in a static context.

It is important to point out that the complexity in word problems (and mathematics tasks in general) is rarely linked to dynamics and kinematic questions, which means that conventional animations might provide limited benefits in this domain. The complexity in word problems rather arises from the multistage solution process and from interactions between the situational and mathematical mental models. Here, a different kind of dynamic visualization might be needed that guides students' attention through the task (Ehrhart & Lindner, 2023) so that these components can be optimally integrated into a mental model.

One recommended way to guide students' attention is to segment the content into smaller pieces and to present these segments over time (e.g., De Koning et al., 2011; Mayer & Moreno, 2003; Ploetzner & Lowe, 2012; van Gog et al., 2010). In fact, segmentation is also used by students as a mental heuristic to reduce the complexity of the task (Verschaffel et al., 2020). Visual segmentations can further reduce the extraneous cognitive load that a visualization evokes (e.g., Mayer & Moreno, 2003); thus, an adequate segmentation might play an important role in word problems (Thevenot & Oakhill, 2005). Using segmentations that include brief pauses between the segments might help to limit students' need of attending to new information while the processing of the previous information in working memory is still ongoing (see, e.g., Mayer & Moreno, 2003; Spanjers et al., 2012). Furthermore, segmentations could highlight the importance of specific parts of the word problem at certain time points (i.e., signaling effect; see, e.g., Mayer & Moreno, 2003; Spanjers et al., 2010, 2012).

So far, there is a lack of empirical research within the field of educational testing (i.e., where existing knowledge needs to be applied to problem-solving situations). Segmentations, for example, have mainly been used to improve learning outcomes and in the context of dynamic systems so far (e.g., Lowe & Boucheix, 2016). A first study shows that using segmented representational visualizations (i.e., dynamic RPs) instead of unsegmented, static representational visualizations may also be supportive in educational testing with word problems (Ehrhart & Lindner, 2023). However, further research is needed (a) to better understand the reliability of dynamic RPs to evoke cognitive and affective-motivational improvements compared to text-only items and (b) to investigate the added value of dynamic RPs compared to static RPs (e.g., to replicate that dynamic RPs provide further cognitive improvements).

DPs in Testing

A second common type of pictures are DPs that visualize information that is not relevant to solve the actual problem (e.g., Gagatsis & Elia, 2004; Lindner, 2020, 2021). Instead, DPs often include interesting but irrelevant aspects (i.e., seductive details) to visualize the broader topic of a task. From a theoretical standpoint, DPs may enhance a task's visual appeal and thereby enhance students' situational interest, their affect, and their test-taking motivation (see, e.g., Lindner, 2020; Magner et al., 2014). This aligns with the idea that the esthetic appeal of DPs in learning materials may improve students' affective state (see Hekkert, 2006; Lenzner et al., 2013). Nevertheless, task-irrelevant details can also occupy students' limited cognitive resources and disrupt the mental modeling process (e.g., Lindner, 2020; Mayer, 2014; Mayer & Moreno, 2003). Accordingly, results in the context of learning are mixed (e.g., see Brom et al., 2018; Sundararajan & Adesope, 2020).

All in all, not much is known about the effects of static DPs in problem-solving tasks (see Hu et al., 2021). The few studies that focused on the use of DPs in test items indicate that conventional static DPs neither improved performance (Berends & van Lieshout, 2009; Lindner, 2020) nor metacognition nor affective-motivational outcomes when compared to text-only items (Lindner, 2020). As a result, the cognitive and affective-motivational effects of adding DPs to test items were shown to be inferior compared to the effects of RPs as well (Lindner, 2020). Given the limited data on static DPs in the field of testing, it is important to provide further information on whether static DPs indeed provide no cognitive and metacognitive improvements (i.e., in line with expectations of test designers), and also no affective-motivational benefits (i.e., contrary to expectations of test designers).

A lack of affective-motivational effects of static DPs may result from different reasons: (a) the visual appeal of DPs in general may not be powerful enough to please students and evoke affective-motivational effects; thus, any type of DP would proof to be ineffective or (b) static DPs in particular may not be sufficient to decorate a word problem that includes multiple content segments. In the second case, (only) static DPs would be ineffective, whereas an optimized version of DPs may evoke affective-motivational benefits. This raises the question of how an optimized version of a DP for word problems might look like.

Dynamic DPs in Testing

Previous meta-analyses in the field of learning found a small positive effect of decorative content on affective-motivational parameters (Brom et al., 2018). The specific design of the seductive content, however, seems to have an important impact on its effectiveness: Negative effects of seductive details were, for example, only found for static seductive details but not for dynamic seductive details (Sundararajan & Adesope, 2020). The design possibilities that arise from the current digitalization allow to implement DPs with a certain kind of dynamic into test items, which might better achieve the anticipated affective-motivational effects and could diversify the presentation of word problems in digital tests.

Previous experiments on dynamic decorative visualizations mainly stem from the area of multimedia learning and target decorative animations. Höffler and Leutner (2007), for example, showed that decorative animations did not have any impact on cognitive outcomes and that their motivational effectiveness seems limited. The term "decorative animations" can be understood in several ways, ranging from video clips with decorative purposes to single decorative elements such as animated GIFs (e.g., Pink & Newton, 2020). Although decorative animations might extend extraneous cognitive load by providing more seductive details in a rapidly changing way (see Höffler & Leutner, 2007), Sundararajan and Adesope (2020) backed the assumption that decorative animations might be less detrimental to students' abilities than static DPs because the seductive details in dynamic DPs are visible only momentarily and, thus, are less attention-grabbing than those in static DPs (see also Castro-Alonso et al., 2018). Although more research on this assumption is needed, it could point to the potential benefits of dynamic DPs in tests (at least in comparison to static DPs), for example, when students perceive cognitive fatigue and lack cognitive resources because of high cognitive demands (see Ackerman & Kanfer, 2009).

However, research on multimedia testing is still at an early stage and the effects of DPs that include a certain kind of dynamic are yet to be explored. Following the idea that conventional animations might not be suitable for word problems as there is not necessarily a process or motion involved (see Ploetzner et al., 2020), it seems again important to avoid the term "animation" and to design a dynamic alternative to static DPs that is more appropriate for word problems. Compared to complex animations, the dynamic of DPs can be created even simpler, by showing a slideshow of separate DPs that follow one another (i.e., dynamic DPs). Such dynamic DPs might be more suitable to decorate the multiple segments and, thus, the entire storyline of a word problem. In this way, dynamic DPs could make the task become more vivid and could increase students' satisfaction, presuming that the students actually concentrate on them and do not ignore the decorative content (Strobel et al., 2019; see also Sanchez & Wiley, 2006). Investigating the effects of dynamic DPs might help to check whether static DPs might not be optimally designed to provide the anticipated affective-motivational improvements. In other words, dynamic DPs might provide an added value compared to text-only items and static DP items, which would result in an actual use case to implement dynamic DPs in tests.

The Present Study

The present preregistered online experiment focused on the cognitive and affective-motivational effects of RPs and DPs in the context of mathematical word problems, while both picture types were presented either in a static or dynamic way. Regarding static RPs, we aimed to replicate and test whether their several supportive effects also hold within an experimental online setting. Dynamic RPs that further segmented the task and visualized the same content as static RPs in smaller pieces over time (see Figure 1) were expected to further reduce the students' mental modeling demands. Regarding static DP, which does not visualize essential task content and only decorates the overall topic of a word problem, we expected to replicate the finding that there were no beneficial effects on item-solving satisfaction. Dynamic DPs, however, that provided a slideshow of multiple DPs one after another (see Figure 2) were expected to eventually evoke beneficial affective-motivational effects by decorating various segments of a word problem. In a direct comparison, the experiment aimed to provide further evidence that RPs are more supportive than DPs, which has previously been shown for static pictures but not yet in the context of dynamic pictures. Our preregistration of all hypotheses and the study design can be found by following this https://aspredicted.org/GF9_T5J.

Hypothesis 1a: Effects of Static RPs (Replication): In line with previous results on the multimedia effect in testing (e.g., Cooper et al., 2018; Lindner, 2020; Lindner et al., 2018; Lindner, Eitel, et al., 2017; Lindner, Lüdtke, et al., 2017; Saß et al., 2012) and expectations based on the expectancy-value theory that cognitive facilitations might evoke a higher expectancy of success (see, e.g., Lindner, 2020), we expected to find an increase in students' performance, metacognitive ratings, and item-solving satisfaction in items with a static RP compared to text-only items. In line with earlier findings (Berends & van Lieshout, 2009; Lindner, 2020), the processing of additional static RPs was expected to extend time-on-task compared to text-only items.

Hypothesis 1b: Effects of Dynamic RPs (Replication): In line with earlier findings (Ehrhart & Lindner, 2023), we expected to find positive effects for items with dynamic RPs compared to text-only items on performance, metacognition, and satisfaction (see also Hypothesis 1a), because the dynamic RPs conveyed the same visual information as the static RPs; the only difference was the temporal segmentation.

Hypothesis 1c: Dynamic Versus Static RPs: With regard to the facilitating effects of segmentations in the context of multistage problems (e.g., Lowe & Boucheix, 2016), we expected items with dynamic RPs to outperform items with static RPs with regard to their effects on performance, metacognition, and satisfaction. However, the temporal segmentation in items with dynamic RPs was expected to extend the students' time-on-task compared to items with static RPs.

Hypothesis 2a: Effects of Static DPs (Replication): Due to the irrelevance of static DPs for the item solution process (i.e., providing no task-related information) and in accordance with previous research findings (see Berends & van Lieshout, 2009; Lindner, 2020), we did not expect to find any significant effects by adding static DPs on students' performance, metacognition, and satisfaction when compared to text-only items.

Hypothesis 2b: Effects of Dynamic DPs: Due to their ability to decorate multiple segments of the word problems, we expected to find positive effects of items with dynamic DPs on students' satisfaction compared to text-only items. However, the higher amount of decorative content in dynamic DPs could also affect students negatively, because of seductive details, so that reduced performance and lower metacognitive ratings could occur, as well as an increased time-on-task, in items with dynamic DPs compared to text-only items.

Hypothesis 2c: Dynamic Versus Static DPs: In accordance with Hypothesis 2a and Hypothesis 2b, for items with dynamic DPs compared to items with static DPs, we expected to find reduced performance and metacognitive ratings but increased satisfaction. Moreover, the segmentation and stepwise changes in dynamic DPs were expected to extend time-on-task compared to static DPs.

Hypothesis 3a: Static RPs Versus Static DPs (Replication): In accordance with Hypothesis 1a and Hypothesis 2a, we expected to replicate the findings that items with static RPs increase performance, metacognitive ratings, and satisfaction more than items with static DPs (Lindner, 2020).

Hypothesis 3b: Dynamic RPs Versus Dynamic DPs: In accordance with Hypothesis 1b and Hypothesis 2b, we expected that items with dynamic RPs would increase performance, metacognition, and satisfaction more than items with dynamic DPs. Moreover, because of the lack of representational information in DPs, we expected to find extended time-on-task in items with dynamic DPs compared to items with dynamic RPs.

Exploratory Analysis

We further conducted exploratory analyses, taking the item position into account to understand how time-on-task effects developed across testing time. We also considered the actual response correctness to investigate whether higher metacognitive ratings corresponded with correct answers but not incorrect answers.

Method

Participants and Study Design

A total of $N = 345$ students[1] at German secondary schools (Grades 5–7) participated in this online experiment from November 2021 to May 2022. All students who quit the study before answering at least 12 of the 18-word problems ($n = 15$), who showed disengaged test-taking behavior ($n = 19$), or who perceived technical difficulties ($n = 3$) were excluded from the analyses. The final sample included 308 students ($M_{age} = 11.63$, $SD_{age} = 1.03$) and encompassed (a) 51% male, 48% female, and 1% diverse students;

[1] Unfortunately, the COVID-19 pandemic had an impact on the experimental phase of this study in terms of (a) a limited amount of schools that allowed us to conduct the study and (b) limited participation rates at each school. As a result, we were not able to reach the preregistered sample size of at least 400 students.

(b) 52% from academic-track and 48% from nonacademic-track schools; and (c) 33% fifth graders, 27% sixth graders, and 40% seventh graders.[2]

In a 3 × 2 mixed design, 18 computer-based mathematical word problems were presented to each student with three different visualization types as a within-subject factor (text-only vs. DPs vs. RPs). Additionally, the dynamics of the visualizations were manipulated as a between-subject factor (static vs. dynamic). To distribute the test items equally across the multimedia conditions, the test items were grouped into three item sets of six items each (i.e., Three Sets × Six Items). We made sure that each item set consisted of word problems with different topics, calculation types, and calculation complexity. The combination of the item sets with different visualization conditions, based on a Latin-square design, resulted in a multimatrix design with three booklets (see Figure 3). The order of the items was always randomized to avoid item position effects and further reduce error variance. With regard to the booklet number and the presence of dynamics, all students were randomly assigned to an experimental condition.

Materials

Testing Material

The mathematical word problems that were used for the present study were inspired by items from established assessments such as the trends in international mathematics and science study (Mullis et al., 2009) and the national educational panel study (e.g., Hahn et al., 2013). The items were created for a specific target group (Grades 5–7), checked for age-appropriate wordings and approved by an expert in mathematics didactics. An exemplary word problem is shown in Figure 1. Many of the items have also already been used successfully in previous studies (e.g., Ehrhart & Lindner, 2023).

The structure of the tasks was always similar: Within each word problem, one or more protagonists performed a certain action which was embedded in an everyday problem-solving situation (e.g., a tour with a rowing boat). Over the course of the task, several pieces of information with a high element interactivity with each other were introduced (e.g., amount of time needed to row different parts of the rowing course). Thus, the students needed to apply their existing declarative and procedural knowledge in a multistage calculation process, which was based on various arithmetic operations in several calculation steps to find the right solution (e.g., how long did the whole tour take). Each item provided four answer options (i.e., multiple-choice format) from which the students had to select the one correct answer option (i.e., single-choice response). The present study was designed as an online experiment. Each word problem included a text-based item stem at the top of the page and a one-sentence question with four answer options underneath. The multiple-choice format was chosen so that students could answer the items by using a mouse cursor or touchpad. As we could not ensure equivalent devices for the participants in the online study setting, we aimed to avoid the need of a physical or virtual keyboard to answer the items. Furthermore, providing items in a multiple-choice format is common in the field of multimedia testing to avoid typing speed to impact time-on-task measures. As shown in Figures 1 and 2, some tasks presented an additional visualization that was displayed between the item stem and the question, depending on the experimental condition.

Manipulation of the Visualizations and Dynamics

The design of the visualizations was dependent on the picture type (i.e., representational vs. decorative) and the kind of dynamic (i.e., static vs. dynamic) so that both picture types were presented in either a static or a dynamic version. In addition to a comparison of both kinds of RPs (see Figure 1) and both kinds of DPs (see Figure 2), we provide a video-based example of all the experimental conditions for an exemplary item in our digital online materials (see https://osf.io/z746n/?view_only=a28044d443ce443bb496ed97a7ac0438). Moreover, the online materials also contains a direct comparison of the three visualization types in both the static and the dynamic conditions.

In the static RP condition, a gray-shaded schematic image with a rather simple abstraction quality was presented, including extra pictorial features (e.g., arrows, brackets) to indicate relations between the elements. Each static RP represented exactly the same information as that in the verbal item stem (i.e., multiple representations), and did not provide any additional information. A schematic presentation with a rather simple abstraction quality was chosen because the mental models that are constructed on the basis of such presentations can support students in applying their analytical and conceptual abilities (see, e.g., Ploetzner et al., 2020); they also follow a cost-efficient design procedure and, thus, seem to be recommended for dynamic adaptions (Höffler & Leutner, 2007; Ploetzner et al., 2020).

The dynamic RP condition was created by making a temporal-dynamic segmentation of the corresponding static RPs. For this purpose, the item stem was subdivided into several thematic segments and the pictorial elements of the static RP were assigned to these segments. In the dynamic RP condition, the different segments were faded in one after another to complement the picture over time. The segments always appeared chronologically, based on their order in the corresponding text. The dynamic RPs started automatically when students entered an item page and students were given enough time to explore the elements of each segment before additional content was faded in (mostly around 2–3 s), as recommended in the literature (Mayer & Moreno, 2003; Moreno, 2007). After all segments were faded in, the final dynamic RP was visually identical to the corresponding static RP and remained visible until an answer was selected. As the dynamic RPs provided the same visual information as the static RPs, they also did not provide any additional information beyond the verbal information in the item stem. Consequently, differences between the static and dynamic RP conditions resulted only from the temporal dynamics. Current design guidelines on how to avoid confounding effects in the comparison of static and dynamic visualizations were followed (see Castro-Alonso et al., 2016). A digital example of the dynamic RPs is provided in the online materials (see https://osf.io/z746n/?view_only=a28044d443ce443bb496ed97a7ac0438) and a direct comparison of the static and dynamic RPs is provided in Figure 1.

In the static DP condition, only one black-and-white picture in photorealistic quality was displayed, which was visible until an answer was selected. This DP visualized the overall topic of the

[2] Please note that in Germany, the Grades 5–7 are usually part of the secondary school system. At the end of Grade 4, students change from elementary school to an academic or nonacademic secondary school track.

Figure 3

Presentation of the Booklets That Distributed the Items Across the Three Visualization Types Within Subjects (i.e., Text-Only, Decorative, and Representational), Following a Latin-Square Design

	Word Problems (18 Items)		
	Item Set 1 (6 Items)	Item Set 2 (6 Items)	Item Set 3 (6 Items)
Booklet 1	Text-only	Decorative	Representational
Booklet 2	Decorative	Representational	Text-only
Booklet 3	Representational	Text-only	Decorative

task but did not provide any details regarding the specific problem-solving situation.

The dynamic DP condition was based on the same thematic segments of the item stem that had already been determined to segment the dynamic RPs. For each segment, an individual static DP was chosen that represented the broader theme of each segment (e.g., Segment 1: picture of a pier, Segment 2: picture of a rowing boat, etc.). In the final dynamic DP, all of these static DPs were presented one after another and appeared chronologically in a slideshow, based on their order in the corresponding text, and at exactly the same time as in the dynamic RPs. As was the case for the dynamic RPs, the dynamic DPs started automatically when students entered a new item page and students were given enough time (see above) to explore each segment before additional content was faded in. When the last DP was presented, the final picture remained visible until an answer was selected. Consequently, the dynamic DPs can be seen as a dynamic slideshow. The use of photo-realistic DPs, which makes it possible to use stock footage, was assumed to be more cost-efficient than designing individual schematic pictures for each segment (i.e., more valid for task designers and teachers) and also seemed appropriate for dynamic contexts (see, e.g., Höffler & Leutner, 2007). Further, we decided to build a slideshow to preserve the decorative nature of dynamic DPs. Slideshows never visualized the information from multiple segments together, so they were less likely to gain representational characteristics over time. A digital example of the dynamic DPs is provided in our online materials (see https://osf.io/z746n/?view_only=a28044d443ce443bb496ed97a7ac0438), and a direct comparison of the static and dynamic DPs is provided in Figure 2.

Measures

Response Correctness

After the students answered the word problems, the correctness of each response was coded dichotomously (i.e., 0 = *incorrect*, 1 = *correct*).

Metacognitive Ratings

The students rated their perceived task difficulty and confidence about solving the item correctly as a metacognitive rating right after solving each word problem. The ratings, based on the measures from Lindner (2020) and Lindner et al. (2022), were assessed with two self-report items (“Working on this task was easy for me”; “I think I have solved this task correctly”) and were rated on a 4-point Likert scale (Cronbach’s $\alpha = .93$). The mean of the two items was used for the analyses. This assessment might also provide tentative evidence regarding students’ extraneous cognitive load (see Sweller, 2010; Sweller et al., 2011) while solving test items with different picture types, but it needs to be interpreted with caution in this regard.

Item-Solving Satisfaction

The students also rated their item-solving satisfaction right after solving each word problem. The ratings, based on the measures from Lindner (2020) and Ehrhart and Lindner (2023), were repeatedly assessed with two self-report items (“Working on this task was fun for me”; “I liked this task”) and were rated on a 4-point Likert scale (Cronbach’s $\alpha = .86$). The mean of the two items was used for the analyses.

Time-on-Task

Students’ time-on-task was measured individually for each word problem from item onset to response selection. To check for extreme response times, the data were analyzed in two different ways: (a) Regarding unrealistically high response times, the data were log-transformed and response times higher than two standard deviations above the respective item mean were replaced with the value of two *SD*s above the item mean. This led to replacements in 1.25% of the response times. (b) Regarding unrealistically low response times (rapid-guessing behavior), the data were tested following the NT15 threshold approach (see Wise & Ma, 2012). Response times below 15% of the average response time for each item, up to a maximum of 10 s, were classified as rapid-guessing behavior (e.g., when a student submitted a response after only 5 s while the average response time across students was 62 s). In line with our preregistration, we classified students who showed rapid-guessing behavior in more than 25% of the items as disengaged and excluded all trials of those participants ($n = 19$) from the analyses in favor of data validity (see Participants and Study Design section). For both adjustment types, the cutoff values were calculated separately for different items and experimental conditions to prevent systematic bias because of the experimental manipulation.

Item Position

For each student, the individual position of a certain test item was logged. The positions of the items were grouped into three categories for the purpose of exploratory analyses: (a) Items 1–6, (b) Items 7–12, and (c) Items 13–18.

Procedure

In advance, an institutional ethics review board approved the present study with regard to the study design and the design of the materials but also the acquisition process and the data management. The study was conducted as an online experiment on computers and laptops. All schools participated voluntarily in the study. Students and their legal guardians were informed beforehand about the study and gave their written consent to participate. In each class, the participation was coordinated individually by the students' mathematics teachers; this procedure allowed for a participation during school lessons or as a homework assignment. The students were able to access the study at a personally chosen time point. At the beginning of the study, students were asked to adapt their browser settings to ensure the correct presentation of the content. This was checked by a test video in the next step. After the students had been welcomed, they answered short questionnaires before then working on to the actual experimental word problems. Instructional videos asked students to create a quiet environment during their participation in the study and informed them about the test item characteristics, such as the single-choice format and the necessity to work accurately but quickly and to make educated guesses in cases of uncertainty. Participation in the entire online study took about 29 min on average.

Transparency and Openness

Our research and article follow the standards of the seventh edition of the American Psychological Association (2020). The procedures and materials of the study were ethically approved by an institutional review board. The hypotheses and the study design were preregistered and can be found by following this https://aspredicted.org/GF9_T5J. Data are available upon request from the first author. Examples of the research materials that were used for the present study are provided in a permanent online repository and can be found by following this https://osf.io/z746n/?view_only=a28044d443ce443bb496ed97a7ac0438.

Analyses

According to the experimental design, each student worked on multiple items and each item was answered by multiple students so that both factors, the participant level (i.e., across items) and the item level (i.e., across participants) could have had an impact on the outcomes. To analyze this cross-classified structure, (generalized) linear mixed-effects models ([G]LMMs) were used (see, e.g., Snijders & Bosker, 2012). Compared to traditional repeated-measures analyses of variance (ANOVAs), these (G)LMMs are preferable as they consider not only the experimental manipulations (i.e., fixed effects) but also unsystematic variance caused by items and students (i.e., random effects; Brauer & Curtin, 2018).

In the first step, dummy-coded models with different random slope structures were computed (Barr et al., 2013). These models always included participants and items as random intercepts but systematically varied the picture type (i.e., RP vs. DP) and the picture dynamics (i.e., static vs. dynamic) as potential random slopes. For each dependent variable, these models were compared based on likelihood ratio tests to detect the random slope structure that is most beneficial, yet economical. For each outcome variable, the final model with the most beneficial random slope structure based on likelihood ratio tests is reported. This resulted in one final dummy-coded model per dependent variable. The fixed effects γ_i of these models can be interpreted similarly to regression coefficients (i.e., focusing on one predictor while the other is fixed at zero).

In a second step, effect-coded models were created, based on the recommended random slope structure, to compute the contrasts and test our hypotheses. Overall, this procedure follows a more conservative approach than traditional repeated-measures ANOVAs and reduces the risk of Type I errors (Barr et al., 2013). To assess the effects of dynamic visualizations, we ensured that the students watched the entire visualization. Thus, in line with our preregistration, we excluded trials from the analyses in which the students submitted a response before the dynamic visualization presentation was finished (which was the case in 2.54% of all trials).[3] The degrees of freedom in our (G)LMMs were calculated based on Kenward–Roger approximations (McNeish, 2017). In addition, because traditional effect size measures such as η^2 for repeated-measures ANOVA do not take the full random-effects structure of the (G)LMMs into account, we computed the marginal and conditional R^2 for each model following Nakagawa et al. (2017). To perform the analyses, the software R 4.0.2 (R Core Team, 2020) with the packages "car" (Fox & Weisberg, 2019), "emmeans" (Lenth, 2020), "performance" (Lüdecke et al., 2023), and "lme4" (Bates et al., 2015) was used. The descriptive parameters and effect size estimations were computed using IBM SPSS Statistics 27. Data are available upon request from the first author.

Results

Means and standard deviations of all the experimental conditions and outcome variables are presented in Table 1. The effect size estimations regarding the pairwise comparisons of the visualization conditions are reported in Table 2. The results from the dummy-coded analyses, which include the estimated fixed effects of our experimental manipulations, the estimated variance components, and the estimated R^2 for each dependent variable, are listed in Table 3.

Effects of RPs

Effects of Static RPs

In line with Hypothesis 1a, our (G)LMMs showed higher performance, $\gamma = 0.53$, $z = 4.74$, $p < .001$, $\eta_p^2 = .118$, higher metacognitive ratings, $\gamma = 0.08$, $t(5052) = 3.21$, $p = .001$, $\eta_p^2 = .063$, and higher satisfaction ratings, $\gamma = 0.09$, $t(5053) = 3.41$, $p < .001$, $\eta_p^2 = .067$, in static RP items compared to text-only items. Contrary to our expectation, we did not find a significant effect on time-on-task, $\gamma = -0.35$, $t(5046) = -0.28$, $p = .777$, in static RP items compared to text-only items.

Effects of Dynamic RPs

In line with Hypothesis 1b, our (G)LMMs showed higher performance, $\gamma = 0.39$, $z = 3.15$, $p = .002$, $\eta_p^2 = .037$, higher metacognitive

[3] We also repeated our analyses for the entire data set (i.e., without these exclusions), which confirmed all of the significant and nonsignificant findings reported.

Table 1
Means and Standard Deviations for Each Outcome Variable in Each Experimental Condition, Computed at Student Level (i.e., Aggregated Across Items)

	Static ($n = 162$)			Dynamic ($n = 146$)		
Outcome variable	Text-only *M* (*SD*)	Decorative *M* (*SD*)	Representational *M* (*SD*)	Text-only *M* (*SD*)	Decorative *M* (*SD*)	Representational *M* (*SD*)
Response correctness	0.65 (0.26)	0.65 (0.26)	0.74 (0.23)	0.69 (0.23)	0.68 (0.27)	0.74 (0.25)
Metacognitive ratings	3.30 (0.59)	3.29 (0.59)	3.39 (0.57)	3.30 (0.58)	3.27 (0.62)	3.37 (0.55)
Item-solving satisfaction	3.03 (0.67)	3.05 (0.65)	3.12 (0.67)	3.11 (0.65)	3.10 (0.67)	3.21 (0.64)
Time-on-task	57.84 (23.46)	58.14 (21.86)	57.42 (22.00)	59.58 (24.96)	59.34 (22.64)	60.68 (21.00)

ratings, $\gamma = 0.09$, $t(5053) = 3.36$, $p < .001$, $\eta_p^2 = .058$, and higher satisfaction ratings, $\gamma = 0.10$, $t(5053) = 3.49$, $p < .001$, $\eta_p^2 = .062$, in dynamic RP items compared to text-only items. Contrary to our expectation, we did not find a significant effect of dynamic RPs on time-on-task, $\gamma = 2.53$, $t(5049) = 1.88$, $p = .060$, compared to text-only.

Dynamic Versus Static RPs

In contrast to Hypothesis 1c, our (G)LMMs did not find a significant difference for performance, $\gamma = 0.12$, $z = 0.68$, $p = .50$, metacognitive ratings, $\gamma = 0.02$, $t(385) = 0.03$, $p = .976$, or satisfaction ratings, $\gamma = 0.09$, $t(371) = 1.25$, $p = .210$, in dynamic RP items compared to static RP items. Moreover, there was no significant effect on time-on-task, $\gamma = 4.91$, $t(445) = 1.91$, $p = .057$, in dynamic RP items compared to static RP items.

Effects of DPs

Effects of Static DPs

In line with Hypothesis 2a, our (G)LMMs did not show any significant differences for performance, $\gamma = 0$, $z = -0.04$, $p = .970$, metacognitive ratings, $\gamma = -0.02$, $t(5052) = -0.86$, $p = .389$, or satisfaction ratings, $\gamma = 0.02$, $t(5053) = 0.67$, $p = .500$, in static DP items compared to text-only items.

Effects of Dynamic DPs

Contrary to Hypothesis 2b, our (G)LMMs did not show any significant differences for performance, $\gamma = -0.04$, $z = -0.31$, $p = .756$, and metacognitive ratings, $\gamma = 0.00$, $t(5057) = 0.07$, $p = .950$, in dynamic DP items compared to text-only items. Moreover, there was also no significant effect on students' satisfaction ratings, $\gamma = 0.00$, $t(5053) = 0.01$, $p = .991$, in dynamic DP items compared to text-only items. Lastly, we did not find a significant effect on time-on-task, $\gamma = 2.56$, $t(5057) = 1.88$, $p = .060$, in dynamic DP items compared to text-only items, indicating that the presence of dynamic DPs did not extend the time-on-task. Yet, additional exploratory analyses that further took the item position into account (see Figure 4) showed an increased time-on-task for items with dynamic DPs versus text-only items for the earlier positions (1–6) in the experiment, $\gamma = 4.78$, $t(5080) = 2.21$, $p = .027$, $\eta_p^2 = .006$. The results also show that the time-on-task in dynamic DP items decreased over the course of the experiment, contrary to the trend in the other dynamic conditions, so that there was no longer a significant difference between items with dynamic DPs and text-only items in the later item positions (i.e., Items 13–18) of the experiment, $\gamma = 0.21$, $t(5086) = 0.10$, $p = .924$. Thus, the time-on-task for dynamic DP items and text-only items was nearly identical at the end. Figure 4 visualizes the estimated parameters for this comparison.

Table 2
Partial η^2 Effect Size Parameters of Each Visualization Condition (Text-Only vs. Decorative vs. Representational) for Both Kinds of Dynamic (Static vs. Dynamic), Computed From Repeated-Measures ANOVAs at Student Level (i.e., Aggregated Across Items)

	Static ($n = 162$)		Dynamic ($n = 146$)	
Visualization condition	Decorative	Representational	Decorative	Representational
Response correctness				
Text-only	0.000	0.118	0.003	0.037
Decorative	—	0.120	—	0.055
Metacognitive ratings				
Text-only	0.003	0.063	0.005	0.058
Decorative	—	0.078	—	0.073
Item-solving satisfaction				
Text-only	0.004	0.067	0.001	0.062
Decorative	—	0.032	—	0.075
Time-on-task				
Text-only	0.000	0.001	0.000	0.004
Decorative	—	0.002	—	0.010

Note. ANOVA = analysis of variance.

Table 3
Parameters of the Dummy-Coded [G]LMMs for Each Outcome Variable

Fixed effect	Response correctness$_{\text{(GLMM)}}$				Metacognitive ratings$_{\text{(LMM)}}$					Item-solving satisfaction$_{\text{(LMM)}}$					Time-on-task$_{\text{(LMM)}}$				
	γ	*SE*	*z*	*p*	γ	*SE*	*df*	*t(df)*	*p*	Γ	*SE*	*df*	*t(df)*	*p*	γ	*SE*	*df*	*t(df)*	*p*
Intercept γ_0	0.82	0.20	4.15	**<.001**	3.31	0.05	125	60.52	**<.001**	3.03	0.06	243	53.77	**<.001**	57.71	2.74	45	21.09	**<.001**
RP[a] γ_1	0.53	0.11	4.74	**<.001**	0.08	0.03	5052	3.21	**<.001**	0.09	0.03	5053	3.41	**<.001**	−0.35	1.25	5045	−0.28	.777
DP[b] γ_2	0.00	0.11	−0.04	.970	−0.02	0.03	5052	−0.86	.389	0.02	0.03	5053	0.67	.500	0.58	1.25	5045	0.47	.640
Dynamic[c] γ_3	0.26	0.17	1.55	.121	−0.01	0.07	381	−0.13	.901	0.08	0.07	368	1.14	.257	2.03	2.56	438	0.79	.430
RP × Dynamic γ_4	−0.14	0.17	−0.86	.390	0.01	0.04	5053	0.27	.786	0.01	0.04	5055	0.23	.818	2.88	1.84	5048	1.57	.117
DP × Dynamic γ_5	−0.03	0.16	−0.21	.836	0.02	0.04	5055	0.63	.528	−0.02	0.04	5056	−0.45	.655	1.98	1.85	5052	1.07	.286
Variance components			Est.				Est.					Est.					Est.		
Intercept student-level			1.19				0.28					0.37					377.60		
Intercept item-level			0.47				0.02					0.01					78.79		
Residual			0.36				0.34					0.31					754.61		
R^2_{marginal}			.011				.006					.018					.003		
$R^2_{\text{conditional}}$			.355				.494					.531					.373		

Note. Statistically significant values are printed in bold type. For each outcome variable, the final model with the most beneficial random slope structure based on likelihood ratio tests is reported. [G]LMMs = (generalized) linear mixed-effects models; LMM = linear mixed-effects model; DP = decorative picture; RP = representational picture; *SE* = standard error; Est. = Estimate. Dummy codings.
[a] 0 = text-only, 1 = representational. [b] 0 = text-only, 1 = decorative. [c] 0 = static, 1 = dynamic.

Dynamic Versus Static DPs

Contrary to Hypothesis 2c, our (G)LMMs did not show any significant difference for performance, $\gamma = 0.229$, $z = 1.33$, $p = .183$, metacognitive ratings, $\gamma = 0.02$, $t(389) = 0.24$, $p = .814$, or satisfaction ratings, $\gamma = 0.07$, $t(375) = 0.90$, $p = .371$, in dynamic DP items compared to static DP items. This indicates that dynamic DPs did not negatively impact students' item-solving process at a cognitive level but also did not induce positive affective-motivational effects in comparison to static DPs. Moreover, there was no significant effect on time-on-task, $\gamma = 4.00$, $t(453) = 1.55$, $p = .122$, in dynamic DP items compared to static DP items, indicating that the additional dynamic in dynamic DPs did not extend the time-on-task.

RPs Versus DPs

Static RPs Versus Static DPs

In line with Hypothesis 3a, our (G)LMMs showed higher performance, $\gamma = 0.53$, $z = 4.78$, $p < .001$, $\eta^2_p = .120$, higher metacognitive ratings, $\gamma = 0.10$, $t(5051) = 4.07$, $p < .001$, $\eta^2_p = .078$, and higher satisfaction ratings, $\gamma = 0.07$, $t(5053) = 2.74$, $p = .006$, $\eta^2_p = .032$, in static RP items compared to static DP items. Contrary to our expectation, there was no significant effect on time-on-task, $\gamma = -0.94$, $t(5046) = -0.75$, $p = .453$, in static RP items compared to static DP items, indicating that static RPs did not extend the time-on-task compared to static DPs.

Dynamic RPs Versus Dynamic DPs

In line with Hypothesis 3b, our (G)LMMs showed higher performance, $\gamma = 0.43$, $z = 3.36$, $p < .001$, $\eta^2_p = .055$, higher metacognitive ratings, $\gamma = 0.09$, $t(5054) = 3.20$, $p = .001$, $\eta^2_p = .073$, and higher satisfaction ratings, $\gamma = 0.10$, $t(5053) = 3.38$, $p < .001$, $\eta^2_p = .075$, in dynamic RP items compared to dynamic DP items. Contrary to expectation, there was no significant effect on time-on-task, $\gamma = -0.03$, $t(5052) = -0.03$, $p = .980$, in dynamic RP items compared to dynamic DP items, indicating that dynamic DPs did not extend the time-on-task compared to dynamic RPs.

Exploratory Analyses on Metacognitive Ratings

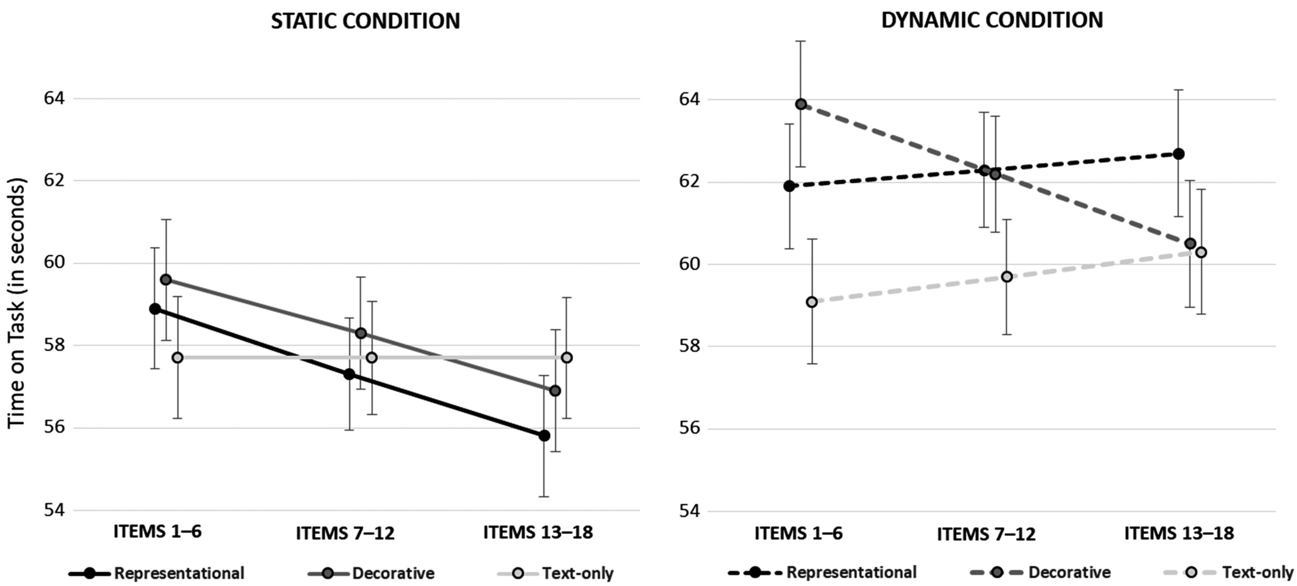
Regarding the effects of static and dynamic RPs on metacognition, additional analyses were performed that also considered whether a student answered an item correctly or incorrectly (i.e., student performance).

For items that were answered incorrectly, our linear mixed-effects model (LMM) detected no significant difference in metacognitive ratings for static RP items compared to text-only items, $\gamma = 0.06$, $t(5069) = 1.29$, $p = .198$, and for static RP items compared to static DP items, $\gamma = 0.09$, $t(5069) = 1.90$, $p = .058$. Also, there were no significant differences in metacognitive ratings for dynamic RP items compared to text-only items, $\gamma = 0.04$, $t(5078) = 0.75$, $p = .452$, and for dynamic RP items compared to dynamic DP items, $\gamma = 0.07$, $t(5078) = 1.34$, $p = .180$.

For items that were answered correctly, our LMM detected no significant difference in metacognitive ratings for static RP items compared to text-only items, $\gamma = 0.06$, $t(5058) = 1.84$, $p = .066$, but significantly higher metacognitive ratings for static RP items compared to static DP items, $\gamma = 0.07$, $t(5058) = 2.43$, $p = .015$.

Figure 4

Estimated Mean Values and Standard Errors From the Effect-Coded Linear Mixed Effects Model on Time-on-Task for Both Kinds of Dynamics (i.e., Static vs. Dynamic; Between-Subjects Factor) and All Visualization Types (i.e., Representational, Decorative, Text-Only; Within-Subjects Factor), Differentiated by the Item Position (i.e., Items 1–6, Items 7–12, and Items 13–18)



Moreover, we found significantly higher metacognitive ratings for dynamic RP items compared to text-only items, $\gamma = 0.09$, $t(5056) = 2.81$, $p = .005$, and for dynamic RP items compared to dynamic DP items, $\gamma = 0.07$, $t(5058) = 2.08$, $p = .037$.

Discussion

In the following, we discuss the empirical, theoretical, and practical conclusions that can be drawn from the results of our experiment, which focused on the cognitive and affective-motivational effects of RPs and DPs, either in a conventional static form or with a dynamic segmentation.

Empirical Contributions

Concerning RPs, an initial empirical contribution of the present study is the replication that students showed higher performance and higher metacognitive ratings in word problems that contained static RPs instead of text only (Ehrhart & Lindner, 2023) and that the multimedia effect in testing could also be demonstrated in an online setting. With the finding of higher satisfaction in the same comparison, this study provides further evidence of affective-motivational gains (e.g., Lindner, 2020) in addition to cognitive gains from the multimedia effect in testing (see Hu et al., 2021; Lindner, 2021). Another contribution is the replication of the finding that the multimedia effect in testing can be extended to RPs that visualize the content dynamically, given that comparable cognitive, metacognitive, and affective-motivational effects were found in items with dynamic RPs (Ehrhart & Lindner, 2023). Apart from the analyses at a global level, our additional exploratory analyses indicate that metacognitive improvements, for example, in items with dynamic RP as compared to text-only items, were specifically reported by the students who actually solved an item successfully. Moreover, we found no significant inflation of metacognitive ratings when students failed to solve an item. This supports the idea that neither static nor dynamic RPs inflated metacognitive ratings beyond proportion and underlines the linking of the cognitive and metacognitive outcomes.

Considering the mixed results in this field, the present findings help to underline the effectiveness of both static and dynamic RPs that act as multiple representations in word problems and do not provide further information besides the item stem. Based on a direct and preregistered comparison, our study provides important evidence that the multimedia effect in testing and related affective-motivational effects may not always be stronger in items with dynamic rather than static RPs, which is in contrast to the findings of more supportive effects of dynamic compared to static RP items in a previous classroom experiment (Ehrhart & Lindner, 2023). The present study further extends the empirical knowledge base on the time-on-task effects of both static and dynamic RPs in word problems. In our study, both types of RPs did not seem to prolong the students' time-on-task when added to the item stem.

Concerning DPs, an important empirical contribution is the finding that both static and dynamic DPs did not have detrimental effects on students' performance but did also not have an impact on their affective state, which was indicated by the absence of cognitive and affective-motivational effects in word problems with either static or dynamic DPs compared to text only. Consequently, the additional dynamic in DPs did not seem to create affective-motivational advantages. However, our explorative analyses indicate that students may have increasingly ignored the dynamic DPs over the course of the experiment when they noticed their irrelevance for solving the word problems. This is based on the finding that students spent more time processing dynamic DPs at the beginning of the test and became quicker over the course of the test. Compared to the decelerating trend that was found in dynamic RP items and text-only items, the dynamic DP items did not have any time difference to text-only items anymore at the end of the experiment. Thus, it stands to reason that the DPs were more or less ignored by the students later on in the course of the test. With regard to the (meta-)cognitive impact of DPs, the present study is in line with previous findings, according to which DPs did not improve performance and did not inflate students' metacognitive ratings inadequately (Lindner, 2020). In line with this, our additional exploratory analyses also indicate that DPs neither improved students' metacognitive ratings when they answered an item incorrectly nor when they answered correctly. Thus, even when focusing on those students who were able to solve the task, DPs did not lead to significant metacognitive improvements; this is in line with the absent cognitive and metacognitive effects of static and dynamic DPs that we found at a global level. Concerning both RPs and DPs, this study was able to replicate the cognitive and affective-motivational advantages of RPs compared to DPs not only for conventional static pictures (see, e.g., Lindner, 2020) but also for dynamic versions of the pictures with a temporal segmentation.

Theoretical Implications

Concerning RPs, the multimedia effect in testing indicates that the presence of RPs in test-based word problems supports the selection and organization of information in both channels (visual and verbal) and facilitates the integration into one coherent mental model, which complies with the underlying multimedia theories (Mayer, 2014; Schnotz & Bannert, 2003). Further, these cognitive facilitations seem to play an important role in word problems, which may be seen as an example for other types of test items with high mental demands, because of the multistage mental processing required (Blum & Leiß, 2007) and the importance of mental modeling skills for solving such tasks (e.g., Fuchs et al., 2006). For static RPs in scientific word problems, eye movement data from Lindner, Eitel, et al. (2017) provided evidence that RPs are especially important for students at the early mental model construction stage of the task (i.e., scaffolding) and at later stages of the problem-solving process (i.e., mental model updating). Although we assume that similar viewing behaviors would be observed in mathematical word problems, this aspect has not been tested in the present study. Further, it seems specifically interesting when and for how long the students actually attended to and make use of the dynamic RPs (i.e., to build a situational model at the beginning of the task; to transfer the situational model in a mathematical model during the task; to retransfer the mathematical model in a situational context at the end of the task).

The results we obtained with regard to students' metacognitive ratings also suggest that students perceived the test items with RPs to be easier, which is consistent with expectations that can be formed based on the CLT (Chandler & Sweller, 1991) and is also in line with students' objective performance increases (Lindner et al., 2022). Moreover, our data suggest an affective impact of RPs on students, indicated by improved satisfaction ratings. However, the causal relationship of effects on different dependent variables was not investigated in this experiment. Based on existing theories, several assumptions can be made regarding the potential relationships. In line with assumptions from the expectancy-value theory (Eccles & Wigfield, 2020) and previous findings in the testing context (Lindner, 2020; Lindner et al., 2018), students' higher satisfaction when working on test items that include RPs may have resulted from lower cognitive effort required to solve the task, causing a higher expectation of success. Also, an improved affective-motivational state, as suggested by the CATLM (Moreno & Mayer, 2007), may have helped students to improve their focus and mental model construction. Beneficial effects on students' affect may also improve the effort that they invest in solving the test items. Apart from that, our explorative finding that metacognitive improvements were exclusively evident for items in which students actually answered correctly is in line with the CATLM assumption that students' performance and their metacognition are mutually linked. However, the causal direction of how the two variables influence each other was not specifically investigated.

Concerning DPs, our findings provide further evidence that the multimedia effect in testing does not result from the mere presence of pictures in word problems and that the inclusion of task-relevant information in a picture is, indeed, essential for improvements in students' mental model construction (Lindner, 2021). This might play an important role in explaining why the usually expected affective-motivational effects of DPs often remain undetectable in practice (see, e.g., Lindner, 2020; Verschaffel et al., 2020). The fact that even more complex dynamic DPs were unable to improve affective-motivational outcomes can may be explained with the modality effect and findings on temporal contiguity from the field of learning (Mayer & Moreno, 2003). Following these explanatory approaches, it might have been difficult for the students to recognize the connection between the two concurrent visual sources, namely the item stem and the decorative slideshow.

On the other hand, the presence of dynamic decorative content could also be expected to increase the mental load compared to static pictures (i.e., extraneous cognitive load; Sweller, 2010). Although we did not measure cognitive load per se, our results on metacognitive ratings suggest that this was probably not the case. Following the findings of studies on transient visual presentations and seductive details (Castro-Alonso et al., 2018; Sundararajan & Adesope, 2020), this might be explained by the idea that decorative information is less attention-grabbing in momentary, transient form than in static form. Furthermore, it is important to consider the potential impact of students' attention control on the effectiveness of static and dynamic DPs, as suggested by previous research on the seductive details effect (Sanchez & Wiley, 2006). It is possible that students deliberately ignored the DPs. Such an intentional ignoring behavior could be influenced by the omnipresence of DPs in today's digital world and may explain the absence of affective and cognitive effects of DPs in the present experiment; an interpretation which is further supported by our time-on-task analyses across testing time.

Practical Implications

The practical conclusions that we draw from the present study focus on the design of mathematical word problems from the perspectives of test designers and test takers.

Concerning RPs, the present study underlines the practical usefulness of both static and dynamic RPs to improve the cognitive processing on the one hand and affective-motivational outcomes on the other hand. Students seem to also recognize the useful character of RPs when they are embedded in online test settings. At the same time, students have no additional costs in terms of time-on-task compared to text-only items, which makes an implementation of RPs even more attractive for test takers. Taking the test designers' perspective into account, the use of both static and dynamic RPs might be an appropriate way to adapt the difficulty of word problems and to improve the student's motivation, which might be especially important in long low-stakes tests in which students might lose their motivation and perceive items as being more demanding (e.g., Lindner et al., 2019; see also Ackerman & Kanfer, 2009). However, the preparation of an individual schematic picture for each task must be considered as a cost that must be paid in order to achieve these effects. As the results of our direct comparison show, static and dynamic RPs turned out as visualizations that were equally effective in the present experiment. This, however, questions the added value of dynamic RPs in the context of word problems, as they pose higher demands on the item creation process.

Concerning DPs, the absence of affective-motivational effects—against common expectations—contradicts the practical implementation of static DPs in test-based word problems. Nonetheless, as they did not have any detrimental impact on cognitive, metacognitive, or affective-motivational outcomes, static DPs could still be used to vary the tasks visually. However, test designers should be aware that these visual variations might not provide any measurable advantages compared to text-only items; therefore, the costs and benefits of designing static DPs should be weighed up carefully. Due to the similar pattern of results found for dynamic DPs, the same recommendations can be made for dynamic DPs, at least in online settings. From a test takers' point of view, working on items with dynamic DPs does not seem to bring any additional costs compared to working on items with static DPs; this also plays an important role in an evaluation of the costs from a test designers' point of view, for example, with regard to the overall test duration. However, because of their limited effectiveness, even the simplified design process required to create dynamic DPs, compared to conventional animations, seems to be disproportionately high, so that a segmentation of DPs may be less advisable for test designers. Taken together, the superiority of RPs over DPs in word problems underlines a clear argument to generally favor static RPs, not only to adapt the cognitive and performance-related outcomes of students but also to evoke positive affective-motivational multimedia effects in test items.

Generalizability and Limitations

The data of this study were gathered in an online experiment, which provided less control compared to classroom settings. Thus, the students may have been easier distracted by external stimuli or may have had a lower level of conscientiousness when working on the study at home. Whereas most of the teachers communicated that they planned to let their students work on the study at home, because of the anonymized data collection, we do not know in which specific setting each student actually participated in the study. Interestingly, however, despite the online context, it was possible to replicate all of the expected effects for static pictures (i.e., static RPs and static DPs), which have been well-researched in classroom studies before. In this regard, the remote context seemed to have had no particular impact on the experimental effects. Still, especially with regard to the novel kinds of dynamic pictures that we implemented here (i.e., dynamic RPs and dynamic DPs), the results should be further investigated for reliability in a classroom setting to check whether context effects may have played a role in the present study.

Future studies should also specifically focus on whether, at what point in time, and for how long the students actually focused on the different visualizations by using eye-tracking methodology (e.g., to better understand scaffolding functions). It could also be interesting to further investigate if RPs can help students to enhance their mental modeling abilities, namely, to perform better in subsequent word problems in terms of a fading approach, even when the items do not include any visuals anymore. Although it was not possible in the present study to acquire a sample size equal to or greater than the preregistered sample size, we are confident that our sample of 308 students in which each student worked on 18-word problems in a mixed design was sufficient to adequately answer our research questions and detect even small effects of interest.

Our results might also depend on the characteristics of the test items, such as the response format (i.e., four predefined answer options) and, most importantly, on the domain (i.e., mathematical word problems). Word problems require a very specific kind of mental model construction, as outlined in the modeling cycle (Blum & Leiß, 2007). Although the segmentation used to create dynamic pictures can be applied to many different kinds of complex test items, the generalizability of the effects to other content domains remains unclear.

Conclusion and Future Directions

Visualizations are an important design element that can help to adapt test items to influence cognitive and affective-motivational experiences in students'. The present experiment investigated how different types of static and dynamic visualizations can impact students' cognitive, metacognitive, and affective-motivational outcomes. In a nutshell, we were able to replicate the cognitive and affective-motivational benefits when adding conventional static RPs to text-only items, whereas dynamic versions of RPs showed similar effects and added no benefits because of the dynamic presentation with visual segmentation. Furthermore, in line with our expectations, we did not find affective-motivational benefits of adding static DPs to text-only items. Yet, beneficial effects were also not present for more complex dynamic DPs that provided multiple static DPs in a slideshow.

Taking the modality effect into consideration, this lack of improvements in the dynamic conditions may have resulted from the fact that two visual sources (i.e., written item stem and visualization) were presented at the same time (e.g., Mayer & Moreno, 2003). At certain stages of the mental model construction, these visual sources might use the same mental channels (Mayer, 2014), which, consequently, increases the cognitive load and might have somewhat negated the potential advantages of (dynamic) visualizations compared to text-only items. To avoid these negative effects and to make optimal use of the two channels, spoken instead of written text is recommended when additional visualizations are presented (i.e., modality effect; Mayer & Moreno, 2003). Consequently, future research should examine how the effects of different picture types (i.e., static vs. dynamic RPs and DPs) develop when these picture types are combined with spoken text and whether the spoken modality might further improve the effectiveness of the dynamic visualizations.

Summarizing the key findings of this study in terms of practical recommendations, test items that use static RPs in mathematical word problems seem to provide the best costs to benefits ratio to adapt both the cognitive and affective-motivational task effects.

The dynamic visualizations did not have any disadvantages as compared to their static counterparts, but they also did not provide any additional advantages. Thus, less complexity seems to be more effective.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163–181. https://doi.org/10.1037/a0015719

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

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Berney, S., & Bétrancourt, M. (2016). Does animation enhance learning? A meta-analysis. *Computers & Education*, *101*, 150–167. https://doi.org/10.1016/j.compedu.2016.06.005

Blum, W., & Leiß, D. (2007). Investigating quality mathematics teaching: The DISUM project. In C. Bergsten & B. Grevholm (Eds.), *Developing and researching quality in mathematics teaching and learning, proceedings of MADIF* (Vol. 5, pp. 3–16). Swedish Society for Research in Mathematics Education.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. https://doi.org/10.1037/met0000159

Brom, C., Stárková, T., & D'Mello, S. K. (2018). How effective is emotional design? A meta-analysis on facial anthropomorphisms and pleasant colors during multimedia learning. *Educational Research Review*, *25*, 100–119. https://doi.org/10.1016/j.edurev.2018.09.004

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

Castro-Alonso, J. C., Ayres, P., & Paas, F. (2016). Comparing apples and oranges? A critical look at research on learning from statics versus animations. *Computers & Education*, *102*, 234–243. https://doi.org/10.1016/j.compedu.2016.09.004

Castro-Alonso, J. C., Ayres, P., Wong, M., & Paas, F. (2018). Learning symbols from permanent and transient visual presentations: Don't overplay the hand. *Computers & Education*, *116*, 1–13. https://doi.org/10.1016/j.compedu.2017.08.011

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

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

De Koning, B. B., Tabbers, H. K., Rikers, R. M., & Paas, F. (2011). Attention cueing in an instructional animation: The role of presentation speed. *Computers in Human Behavior*, *27*(1), 41–45. https://doi.org/10.1016/j.chb.2010.05.010

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, *61*, Article 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

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

Flake, J. K., Barron, K. E., Hulleman, C., McCoach, B. D., & Welsh, M. E. (2015). Measuring cost: The forgotten component of expectancy-value theory. *Contemporary Educational Psychology*, *41*, 232–244. https://doi.org/10.1016/j.cedpsych.2015.03.002

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., Schatschneider, C., & Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, *98*(1), 29–43. https://doi.org/10.1037/0022-0663.98.1.29

Gagatsis, A., & Elia, I. (2004). *The effects of different modes of representation on mathematical problem solving*. Paper presented at the 28th meeting of the international group for the psychology of mathematics education, Bergen, Norway.

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., Dalehefte, I. M., & Prenzel, M. (2013). Assessing scientific literacy over the lifespan—A description of the NEPS science framework and the test development. *Journal for Educational Research Online*, *5*(2), 110–138. https://doi.org/10.25656/01:8427

Harp, S. F., & Mayer, R. E. (1997). The role of interest in learning from scientific text and illustrations: On the distinction between emotional interest and cognitive interest. *Journal of Educational Psychology*, *89*(1), 92–102. https://doi.org/10.1037/0022-0663.89.1.92

Hekkert, P. (2006). Design aesthetics: Principles of pleasure in design. *Psychology Science*, *48*(2), 157–172.

Höffler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction*, *17*(6), 722–738. https://doi.org/10.1016/j.learninstruc.2007.09.013

Hoogland, K., Pepin, B., de Koning, J., Bakker, A., & Gravemeijer, K. (2018). 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

Jitendra, A. K., Griffin, C. C., Deatline-Buchman, A., & Sczesniak, E. (2007). Mathematical word problem solving in third-grade classrooms. *The Journal of Educational Research*, *100*(5), 283–302. https://doi.org/10.3200/JOER.100.5.283-302

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

Leiss, D., Plath, J., & Schwippert, K. (2019). Language and mathematics-key factors influencing the comprehension process in reality-based tasks. *Mathematical Thinking and Learning*, *21*(2), 131–153. https://doi.org/10.1080/10986065.2019.1570835

Lenth, R. (2020). *Emmeans: Estimated marginal means, aka least-squares means*. R package version 1.4.8. https://CRAN.R-project.org/package=emmeans

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

Lindner, M. A. (2020). Representational and decorative pictures in science and mathematics tests: Do they make a difference? *Learning and Instruction*, *68*, Article 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.), *Cambridge handbook of multimedia learning* (3rd ed., pp. 552–565). Cambridge.

Lindner, M. A., Eitel, A., Strobel, B., & Köller, O. (2017). 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., 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., Lüdtke, O., Grund, S., & Köller, O. (2017). 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., 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*, Article 1533. https://doi.org/10.3389/fpsyg.2019.01533

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

Lowe, R. K., & Boucheix, J.-M. (2016). Principled animation design improves comprehension of complex dynamics. *Learning and Instruction*, *45*, 72–84. https://doi.org/10.1016/j.learninstruc.2016.06.005

Lüdecke, D., Makowski, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., Wiernik, B. M., Arel-Bundock, V., Thériault, R., Jullum, M., & Bacher, E. (2023). *performance: Assessment of regression models performance (0.10.3)*. https://CRAN.R-project.org/package=performance

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

Mayer, R. E. (Ed.). (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge. https://doi.org/10.1017/CBO9781139547369

Mayer, R. E., & Hegarty, M. (1996). The process of understanding mathematical problems. In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 29–53). Mahwah: Erlbaum.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research*, *52*(5), 661–670. https://doi.org/10.1080/00273171.2017.1344538

Moreno, R. (2007). Optimising learning from animations by minimising cognitive load: Cognitive and affective consequences of signalling and segmentation methods. *Applied Cognitive Psychology*, *21*(6), 765–781. https://doi.org/10.1002/acp.1348

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

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *Trends in International Mathematics and Science Study (TIMSS). TIMSS 2011*. International Association for the Evaluation of Educational Achievement (IEA).

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), Article 20170213. https://doi.org/10.1098/rsif.2017.0213

Orrantia, J., Múñez, D., & Tarín, J. (2014). Connecting goals and actions during reading: The role of illustrations. *Reading and Writing*, *27*(1), 153–170. https://doi.org/10.1007/s11145-013-9437-4

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. https://doi.org/10.1093/acprof:oso/9780195066661.001.0001

Pink, A., & Newton, P. M. (2020). Decorative animations impair recall and are a source of extraneous cognitive load. *Advances in Physiology Education*, *44*(3), 376–382. https://doi.org/10.1152/advan.00102.2019

Ploetzner, R., Berney, S., & Bétrancourt, M. (2020). A review of learning demands in instructional animations: The educational effectiveness of animations unfolds if the features of change need to be learned. *Journal of Computer Assisted Learning*, *36*(6), 838–860. https://doi.org/10.1111/jcal.12476

Ploetzner, R., & Lowe, R. (2012). A systematic characterisation of expository animations. *Computers in Human Behavior*, *28*(3), 781–794. https://doi.org/10.1016/j.chb.2011.12.001

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.Rproject.org/

Sanchez, C. A., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition*, *34*(2), 344–355. https://doi.org/10.3758/BF03193412

Saß, S., Wittwer, J., Senkbeil, M., & Köller, O. (2012). Pictures in test items: Effects on response time and response correctness. *Applied Cognitive Psychology*, *26*(1), 70–81. https://doi.org/10.1002/acp.1798

Schnotz, W. (2005). Integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 325–337). Cambridge.

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

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Spanjers, I. A. E., van Gog, T., & van Merriënboer, J. J. G. (2010). A theoretical analysis of how segmentation of dynamic visualizations optimizes students' learning. *Educational Psychology Review*, *22*(4), 411–423. https://doi.org/10.1007/s10648-010-9135-6

Spanjers, I. A. E., van Gog, T., Wouters, P., & van Merriënboer, J. J. G. (2012). Explaining the segmentation effect in learning from animations: The role of pausing and temporal cueing. *Computers & Education*, *59*(2), 274–280. https://doi.org/10.1016/j.compedu.2011.12.024

Stephany, S. (2021). The influence of reading comprehension on solving mathematical word problems: A situation model approach. In E. Gürsoy, A. Fritz-Stratmann & M. Herzog (Eds.), *Diversity dimensions in mathematics and language learning. Perspectives on culture, education and multilingualism* (pp. 370–395). De Gruyter Mouton.

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

Sundararajan, N., & Adesope, O. (2020). Keep it coherent: A meta-analysis of the seductive details effect. *Educational Psychology Review*, *32*(3), 707–734. https://doi.org/10.1007/s10648-020-09522-4

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.

Thevenot, C., & Oakhill, J. (2005). The strategic use of alternative representations in arithmetic word problem solving. *The Quarterly Journal of Experimental Psychology Section A*, *58*(7), 1311–1323. https://doi.org/10.1080/02724980443000593

Trendtel, M., & Robitzsch, A. (2018). Modeling item position effects with a Bayesian item response model applied to PISA 2009–2015 data. *Psychological Test and Assessment Modeling*, *60*(2), 241–263.

Van Gog, T., Paas, F., & Sweller, J. (2010). Cognitive load theory: Advances in research on worked examples, animations, and cognitive load measurement. *Educational Psychology Review*, *22*(4), 375–378. https://doi.org/10.1007/s10648-010-9145-4

Verschaffel, L., Schukajlow, S., Star, J., & van Dooren, W. (2020). Word problems in mathematics education: A survey. *ZDM*, *52*(1), 1–16. https://doi.org/10.1007/s11858-020-01130-4

Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at the 2012 annual meeting of the national council on measurement in education, Vancouver, Canada.

Received November 28, 2022
Revision received May 26, 2023
Accepted July 22, 2023 ■