Contemporary Educational Psychology 51 (2017) 482–492

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych

The merits of representational pictures in educational assessment: Evidence for cognitive and motivational effects in a time-on-task analysis

Marlit Annalena Lindner[a,*], Oliver Lüdtke[a,b], Simon Grund[a], Olaf Köller[a]

[a] Leibniz Institute for Science and Mathematics Education (IPN), Kiel, Germany
[b] Centre for International Student Assessment (ZIB), Germany

ARTICLE INFO

Keywords:
Multimedia effect in testing
Representational pictures
Test-taking motivation
Time on task
Rapid-guessing behavior

ABSTRACT

Adding representational pictures (RPs) to text-based items has been shown to improve students' test performance. Focusing on potential explanations for this *multimedia effect in testing*, we propose two functions of RPs in testing, namely, (1) a *cognitive facilitation function* and (2) a *motivational function*. We found empirical support for both functions in this computer-based classroom experiment with $N = 410$ fifth and sixth graders. All students answered 36 manipulated science items that either contained (*text-picture*) or did not contain (*text-only*) an RP that visualized the text information in the item stem. Each student worked on both item types, following a rotated within-subject design. We measured students' (a) solution success, (b) time on task (TOT), and identified (c) rapid-guessing behavior (RGB). We used generalized and linear mixed-effects models to investigate RPs' impact on these outcome parameters and considered students' level of test engagement and item positions as covariates. The results indicate that (1) RPs improved all students' performance across item positions in a comparable manner (*multimedia effect in testing*). (2) RPs have the potential to accelerate item processing (*cognitive facilitation function*). (3) The presence of RPs reduced students' RGB rates to a meaningful extent (*motivational function*). Overall, our data indicate that RPs may promote more reliable test scores, supporting a more valid interpretation of students' achievement levels.

1. Introduction

Various types of pictures can be found in large-scale assessment (LSA) studies. With computerized tests being applied in LSA more and more, multimedia elements will certainly play an even greater role in the future (Bennett et al., 1999; Greiff, Wüstenberg, Holt, Goldhammer, & Funke, 2013; Wirth, 2008) because complex depictive information can be displayed easily (Wu, Kuo, Jen, & Hsu, 2015; Zenisky & Sireci, 2002). In order to use pictorial elements to their full potential when constructing test items, it is essential to understand their impact on psychometric, cognitive, and motivational variables (Jarodzka, Janssen, Kirschner, & Erkens, 2015; Wu et al., 2015).

While there is still a research gap regarding the functions of pictures in testing, in line with extensive multimedia research in the instructional sciences (see Carney & Levin, 2002; Mayer, 2005; Vekiri, 2002 for reviews), there is also evolving evidence that pictures influence item parameters, item processing, and students' motivation in testing. Comparable to the well-known *multimedia effect in learning* (e.g., Ainsworth, 2006; Mayer, 2005), studies have shown that adding a *representational picture* (RP) that visualizes the item-stem text but does not add other solution-relevant information (i.e., *multiple representations*; Ainsworth, 2006; Carney & Levin, 2002) fosters students' test performance (i.e., *multimedia effect in testing*; Hartmann, 2012; Lindner, Eitel, Strobel, & Köller, 2017; Lindner, Ihme, Saß, & Köller, 2016; Saß, Wittwer, Senkbeil, & Köller, 2012). Moreover, RPs improve students' self-reported test-taking pleasure (Lindner et al., 2016), while there is also tentative evidence that graphical elements might enhance students' test-taking behavior (Wise, Pastor, & Kong, 2009). This suggests that RPs have beneficial cognitive *and* affective-motivational effects in testing. However, the interplay between cognitive and motivational effects in particular deserves closer attention. Thus, in this computer-based classroom experiment, we focused on indicators of both the proposed (1) *cognitive function* and the (2) *motivational function* of RPs in testing at the level of students' item-solving behavior.

1.1. Cognitive multimedia theories

So far, no specific theories have addressed the issue of multiple representations in testing. However, recent studies support the claim that the beneficial effects of RPs on students' test performance may be

[*] Corresponding author at: Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany.
E-mail address: mlindner@ipn.uni-kiel.de (M.A. Lindner).

http://dx.doi.org/10.1016/j.cedpsych.2017.09.009

Available online 30 September 2017

explained by transferring cognitive multimedia *learning* theories to the context of *testing* (e.g., Jarodzka et al., 2015; Lindner et al., 2016). This is because both learning and testing require students to encode and understand the given information: In both situations, students have to build a coherent mental model of the information presented (Glenberg & Langston, 1992; Lindner et al., 2016; Mayer, 2005; Saß et al., 2012; Schnotz & Bannert, 2003; Schnotz et al., 2014) because they cannot perform well in either situation without correctly understanding the material. Thus, as a foundation for the subsequent discussion on how RPs may influence cognitive processes in testing, we first outline assumptions from the *Cognitive Theory of Multimedia Learning* (CTML; Mayer, 2005) and the *Integrated Theory of Text and Picture Comprehension* (ITPC; Schnotz & Bannert, 2003), which both focus on processes of text-picture integration in learning.

1.1.1. Text-picture processing in learning

The CTML (Mayer, 2005) assumes that a dual coding of information by text and picture allows students to use both the auditory-verbal and the visual-pictorial channel in working memory (*dual-channel assumption*) and, thus, to better exploit their limited working-memory capacity (*limited capacity assumption*). It also assumes that students process multimedia material more actively (i.e., selecting and organizing relevant words and images; *active processing assumption*) because text-based and image-based representations need to be integrated with prior knowledge in order to construct a coherent mental representation.

According to the ITPC (Schnotz & Bannert, 2003), a situational mental model can be constructed based on a text in a descriptive branch of representations or based on a picture in a depictive branch of representations. These branches represent two separate but connected channels in working memory. Specifically, when processing a text, a mental representation of the text's surface structure needs to be constructed, from which a propositional representation of the semantic content is generated (Van Dijk & Kintsch, 1983). This process involves inferring interrelations (e.g., spatial relations) that are often only implicitly stated in a text (Glenberg & Langston, 1992). This might lead to erroneous interpretations. As a result, the text-based mental model may inadequately reflect the presented content. In contrast to this, pictures are more efficient to process than text because they allow information to be extracted immediately at a perceptional level and with high accuracy. Hence, the picture's visuospatial relations can be mapped onto semantic relations to provide the structure of the mental model (*analogical structure mapping*; Schnotz & Bannert, 2003). Accordingly, processing a picture with a corresponding text can significantly support the construction of a coherent mental model, for example, by disambiguating ambiguous text passages (i.e., *constraining interpretation function*; Ainsworth, 2006). Therefore, pictures can facilitate or sometimes even replace at least part of the mental model construction based on a text. This applies especially to early picture processing because the picture can already provide a structure of the mental model and facilitate subsequent processing of the text (*scaffolding function*; see, e.g., Eitel & Scheiter, 2015 for a review).

1.1.2. Text-picture processing in testing

Transferring assumptions from the CTML and the ITPC to the context of testing, we also assume that a dual coding of the item-stem text information with a corresponding picture better exploits students' working-memory capacity (cf. Mayer, 2005), while providing faster and more elaborate access to important item-stem information (cf. Schnotz & Bannert, 2003). Thus, in testing, RPs could also help students to build a coherent mental model of the presented problem with less effort and may also serve as mental scaffolds that facilitate better understanding of subsequent text information and thereby prevent misinterpretations. In the answering process, the RP could further support students in taking the transforming and reasoning steps necessary to solve the task at hand (cf. Butcher, 2006; Schnotz & Kürschner, 2008; Sweller, van Merriënboer, & Paas, 1998). Likewise, the RP may be helpful in providing a mental model update of the item-stem information in the answering process as the content can be extracted from a picture more easily than by re-reading a text (Schnotz et al., 2014). Considering these cognitive facilitations that RPs may provide, from a cognitive load perspective (Sweller et al., 1998), integrating RPs into test items may also reduce the extraneous load of the testing material.

1.2. Functions of representational pictures in testing

With regard to the findings that RPs foster students' performance (e.g., Lindner et al., 2016; Saß et al., 2012), there may be different explanations for this multimedia effect in testing. Building on earlier research, we propose that RPs have two major functions in assessment: (1) a *cognitive facilitation function* and (2) a *motivational function*. However, before we explain these functions in greater detail, students' test-taking motivation and test engagement need to be considered as important factors for understanding our assumptions regarding RPs' impact on students' test-taking behavior. Furthermore, our assumptions refer to the case in which test items require *controlled processing* (i.e., problem-solving tasks; cf. *Dual Processing Theory*; Shiffrin & Schneider, 1977), in which higher levels of *time on task* (TOT[1]) tend to be related to better test outcomes (Goldhammer et al., 2014).

Students' willingness to engage in solving a test plays a major role in their test-taking behavior and test outcome. Especially in low-stakes assessments (e.g., LSA studies), in which neither good nor bad test results have direct consequences for the participating students, a certain number of students usually have low levels of test-taking motivation (Wise & DeMars, 2005, 2010). Such students put less effort into solving the tasks, which is, for example, indicated by their tendency to show *rapid-guessing behavior* (RGB; e.g., Finn, 2015; Kong, Wise, & Bhola, 2007). RGB is defined as a very fast, effortless response given to an item. Thus, RGB is usually operationalized by extremely low response times in computerized tests (Wise & Kong, 2005). As RGB represents students' unwillingness to solve an item, it constitutes a negative indicator of students' test-taking motivation (Finn, 2015; Wise & DeMars, 2005, 2010; Wise & Kong, 2005). Accordingly, such students spend less TOT solving test items as a reflection of their lack of effort, but not as a reflection of faster cognitive item processing. In light of this, successfully enhancing the motivation of less engaged students to work on a test (e.g., by design factors, such as RPs) should result in reduced rates of RGB, reflecting increased effort and, as a consequence, improved reliability of the responses given.

However, this does not likewise apply to engaged students, who do not show counterproductive test-taking behavior such as RGB and who invest effort throughout a test, in compliance with the task's requirements for successful completion (i.e., investing sufficient TOT to solve each item). Therefore, the TOT of engaged students should primarily reflect the *cognitive processing time* that is actually needed to encode, understand, and solve a specific item, at least at the beginning of the test session when the students are still cognitively fit and fully motivated to work on the test. Thus, in this study, we interpret engaged students' TOT at the test start (i.e., in early item positions) as a valid indicator of RPs' potential *cognitive facilitations*, because motivational effects are expected to play a minor role in this specific constellation. In the following, we explain these assumptions in greater detail.

1.2.1. A cognitive facilitation function

Alongside the central findings of a multimedia effect in testing (e.g., Hartmann, 2012; Lindner et al., 2016; Saß et al., 2012), there is now process-oriented support for transferring the assumptions from the

[1] The usage of this term varies in the literature; here, *time on task* (TOT) refers to the time a student needs to answer a single item from task onset to providing a response (Goldhammer et al., 2014). TOT is thus interchangeable with the term *response time* and provides an indicator of students' overall item processing speed.

CTML and ITPC to the context of testing, provided by an eye-tracking study by Lindner et al. (2017). The study showed that the fixation time students spent on processing RPs in the item stem was compensated for by a smaller amount of fixation time they spent reading the corresponding text. Students' eye movement behavior further revealed that the picture received particular attention right after item onset, reflecting a use of the RP as a mental scaffold (see, e.g., Eitel, Scheiter, Schüler, Nyström, & Holmqvist, 2013). The picture was also used in the later phase of item solving for a mental model updating when students evaluated the answers with regard to the problem described (Schnotz et al., 2014). Moreover, in text-picture items, students spent significantly less time fixating *incorrect* answer options, and this behavior was associated with better test scores in general. Altogether, and reflecting a trade-off between (a) the time spent encoding the added RP and (b) the cognitive advantages gained from processing the RP (e.g., reduced reading time of the text), there was no significant difference between the TOT spent on text-picture and parallel text-only items in the eye-tracking study. A similar finding was reported in a study by Saß et al. (2012), in which adding an RP to the item stem did not significantly affect students' TOT in a computerized test.

However, considering the different TOT interpretations for engaged and less engaged students we discussed above, the cognitive effects of RPs might have been underestimated in both studies, as students' level of motivational engagement was not controlled for. As the expected motivational effects of RPs might influence students' item-solving effort (i.e., increase in TOT) and obscure the expected cognitive facilitation effect (i.e., reduction in TOT), only the net *cognitive processing time* necessary to solve an item (without a motivational bias) provides a valid indicator of RPs' cognitive facilitations. According to this, even though we assume that RPs' cognitive facilitations support all students in a similar way, we do not think that the effect is traceable in all students at all times. Thus, in the present study we used a continuous analysis to focus specifically on the TOT of engaged students at the test start (i.e., the first item positions) in order to reconsider the reported findings of comparable solution times for text-only and text-picture items. This is because the proposed motivational effects of RPs should have a minimal impact on the TOT variable in this constellation. The TOT measure in this subgroup should therefore constitute a more suitable indicator of the net cognitive item processing time, making it possible to evaluate RPs' maximum potential to accelerate item solving.

1.2.2. A motivational function

In multimedia learning, the affective-motivation related functions of pictures have been discussed for several decades (see, e.g., Carney & Levin, 2002; Lenzner, Schnotz, & Müller, 2013; Leutner, 2014; Mayer, 2014; Moreno & Mayer, 2007). Studies show that pictures have the potential to induce better moods, more alertness and calmness, to reduce the perceived difficulty of the material (Lenzner et al., 2013) and to enhance students' situational interest (Magner, Schwonke, Aleven, Popescu, & Renkl, 2014). Situational interest refers to a short-term state component of interest and can be influenced by the specifics of the environmental setting (Hidi, 2006; Hidi & Renninger, 2006; Mitchell, 1993). According to Schiefele (2009), evoked (triggered and maintained) situational interest is an antecedent of intrinsic motivation and involves enjoyment, focused attention, and increased cognitive function (Ainley, Hidi, & Berndorff, 2002; Hidi, 2006; Hidi & Renninger, 2006), and can, as a consequence, improve performance (e.g., Ainley et al., 2002). RPs might trigger students' situational interest and foster their motivation not only in learning, but also in testing situations. Particularly in low-stakes assessment, where students often lack intrinsic motivation to work on a test, it is important to consider design factors that may improve students' test-taking motivation in order to obtain a better measure of their actual achievement, based on the test results (see also Baumert & Demmrich, 2001; Wise & DeMars, 2005; Wise et al., 2009). We expected RPs to positively influence the motivation of less engaged students in particular, and expected this to be reflected in improved test-taking behavior (i.e., less RGB).

The motivational effects of RPs might play an even larger role over the course of testing because students need to stay motivated in order to familiarize themselves with different tasks and topics (i.e., to build mental models) in a short time frame and to cognitively switch between solution strategies. Due to the exertion of effort needed when working on cognitively demanding tasks, it is assumed that such tasks become aversive over the course of the test, and that this leads to an increase in students' fatigue and a progressive reduction in their motivational effort (Ackerman & Kanfer, 2009; Ackerman, Kanfer, Shapiro, Newton, & Beier, 2010; Inzlicht & Schmeichel, 2012). LSA studies show that the probability of successfully solving test items decreases over the course of testing (i.e., identical items are more difficult when presented in later positions); this is known as the *item position effect* (Debeer, Buchholz, Hartig, & Janssen, 2014; Hartig & Buchholz, 2012). Correspondingly, a progressive decrease in students' motivation and persistence is reflected in higher RGB rates (Wise et al., 2009) and a TOT decrease over the course of testing (Penk & Richter, 2016).

Evidence for a motivational function of RPs was provided in an experiment by Lindner et al. (2016) that revealed that students' self-reported test-taking pleasure was higher when solving text-picture compared to text-only items. Moreover, an explorative field study by Wise et al. (2009) found that test items containing some kind of graphical element (e.g., picture, diagram) were associated with lower RGB rates. While Wise et al. (2009) documented an increase in RGB with progressive testing time, they also identified an interaction between the RGB-reducing effect of graphical elements and the item position effect. Thus, graphics seemed to be more effective in preventing RGB rates from increasing when presented in the later phase of the test. However, all of these findings of Wise et al. (2009) must be interpreted carefully because their correlative field study was not deliberately designed[2] to investigate the impact of graphics on students' test-taking behavior (e.g., no control group, only a few items contained graphics). Thus, experimental studies are needed to put the causality of their correlative findings to the test.

2. The current study

The present experiment was designed to investigate the assumed *cognitive* and *motivational* functions of RPs in a classroom setting. Based on the literature and the theories outlined above, we aimed to identify RPs' impact on students' solution success and on their test-taking behavior (i.e., TOT, RGB), taking the level of students' test engagement and item position effects into account. To identify engaged and less engaged students, we interpreted students' number of RGB trials in the test as a negative indicator of students' test engagement (Finn, 2015; Wise & DeMars, 2005, 2010; Wise & Kong, 2005) and used this measure, together with an item position variable, as covariates in our solution-success and TOT analyses. We also considered RPs' impact on students' RGB per item as a dichotomous outcome parameter (i.e., RGB/no RGB), reflecting a lack of motivation and effort to solve specific items in the test (i.e., text-only vs. text-picture items). In conducting this study, we aimed to test the following hypotheses (see Table 1 for an overview):

(1) Multimedia effect in testing hypothesis (H1). We expected to replicate earlier findings that students are more successful (i.e., more likely to give a correct response) when solving text-picture items compared to text-only items. We assumed that this multimedia effect in testing reflects both the cognitive and the motivational effects of RPs. We further explored the interaction between the multimedia effect and

[2] The field study of Wise et al. (2009) explored the correlational relations between several person and item factors that may influence the occurrence of rapid-guessing behavior (RGB), whereby the presence of (unspecified) graphical elements in the items was only one aspect considered, and was not the focus of their research. Thus, neither the use of graphics nor the positions of the graphics in the test were experimentally controlled for.

Table 1
Overview of the hypotheses.

Hypothesis Name		Outcome Parameter	Expected Effects by the Representational Picture (RP) Manipulation
Multimedia Effect in Testing Hypothesis	*H1*	Solution Success	Students are more successful when solving text-picture items compared to text-only items (i.e., multimedia effect in testing); we assume that this reflects both the cognitive and the motivational benefits of RPs.
Cognitive Facilitation Hypothesis	*H2*	Time on Task	Engaged students solve text-picture items faster than (or at least as fast as) text-only items at the beginning of the test (i.e., RPs have the potential to accelerate the item solving process, reflecting a cognitive facilitation effect).
Motivational Function Hypothesis	*H3a*	Rapid-Guessing Behavior	Students less often show rapid-guessing behavior (RGB) when they are confronted with text-picture items compared to text-only items, reflecting a motivation-enhancing effect of RPs.
	H3b		RPs' impact on preventing students from engaging in RGB increases over the course of testing, reflecting a perseverance-enhancing effect of RPs with regard to students' motivation.

students' level of test engagement, as well as the development of the multimedia effect across testing time (i.e., item positions).

(2) Cognitive facilitation hypothesis (H2). Based on earlier findings, we assumed that RPs compensate for the encoding time they require by helping students to build a coherent mental model more effectively and by supporting the answering process. Thus, we expected to find an equal or even reduced TOT in text-picture compared to text-only items. However, we propose that such a cognitive facilitation of RPs is only accurately traceable at the test start (i.e., in the first item positions) and in the group of engaged students, because motivational factors are not expected to influence students' TOT in this specific constellation and thus deliver a more valid indicator of the cognitive processing time that is necessary to solve an item. We further explored the impact of students' level of test engagement and item positions on TOT and their relations to the effect of RPs, in order to ensure a conclusive data analysis.

(3) Motivational function hypothesis (H3). We expected (*H3a*) that students become motivated by RPs to put more effort into solving test items and therefore show less RGB in text-picture compared to text-only items. Because RPs might be especially helpful in keeping students' motivated in the later phase of a test, we further expected (*H3b*) that the impact RPs have on preventing students from showing RGB increases over the course of testing (i.e., item positions), reflecting a motivational perseverance effect.

3. Method

3.1. Sample and study design

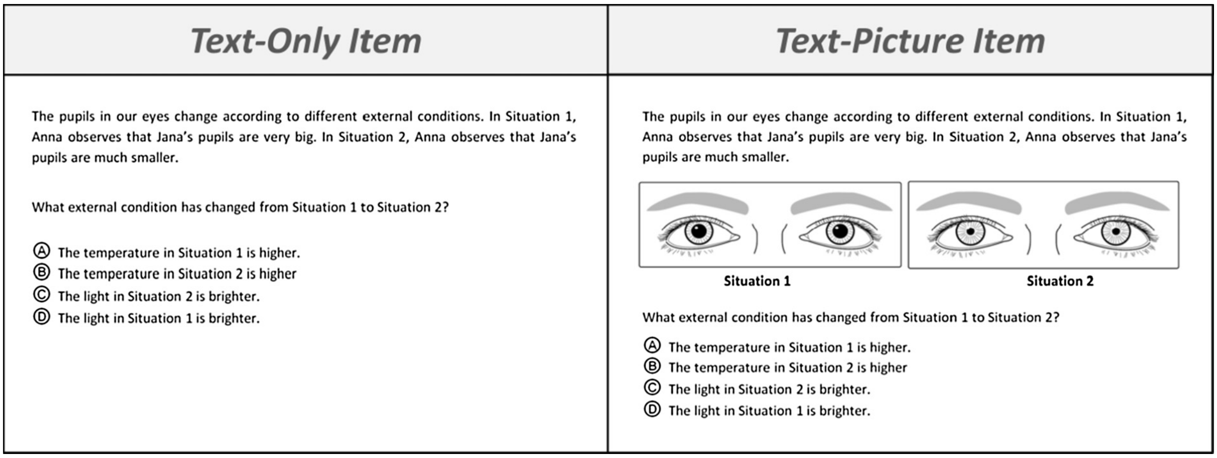
The sample comprised $N = 410$ students in the fifth and sixth grades in three schools in the northern part of Germany. Due to technical problems and other dropout, the data of nine students were unavailable for analysis. Thus, the reported results refer to a final sample of $N = 401$ students (53.4% female, 51.4% fifth grade, $M_{\text{age}} = 10.74$, $SD_{\text{age}} = 0.76$). Of these, $n = 247$ students attended academic track schools (Gymnasium) and $n = 154$ students attended a non-academic track school (regional school). The students were recruited in classes by their school principal or class teacher. All students were informed that their individual participation was completely voluntary and that they would not have to face any negative consequences if they did not participate or if they canceled their participation. Students also had to obtain written permission from their parents or legal guardians before participating.

Our design followed a within-subject manipulation regarding the presentation of the testing material (*text-only* vs. *text-picture*): We grouped text-only items and corresponding text-picture items into three parallel test blocks (12 items per block) and presented them in a balanced multi-matrix design to control for person- and item-related factors that could affect the findings. Each of the resulting six booklets contained 36 items; at least one block was presented as a text-only and one block as a text-picture version. This ensured that each student answered at least 12 items under both experimental conditions while never answering the same item twice. All items were randomly assigned to the test blocks. A randomization check confirmed that the item difficulty did not differ between the blocks, $F(2, 33) = 0.05$; $p = .95$; $\eta^2_p = 0.003$. To investigate the multimedia effect over the course of the test (i.e., item positions), items were presented in a random order for each student within each test block to avoid always presenting certain items in certain positions. The six booklets were randomly assigned to the students and were equally distributed in the sample, including the different school tracks. Each text-only item was solved by at least $n = 191$ students ($M = 196$; $SD = 0.96$) and each text-picture item by at least $n = 197$ ($M = 204$; $SD = 0.81$) students. All students completed the science test. Thus, apart from a minimal data loss (0.1%) due to technical issues, there were no missing values.

3.2. Materials

3.2.1. Testing material

In this study, we used 36 multiple-choice items that were constructed in close relation to the science framework of the *Trends in International Mathematics and Science Study* (TIMSS; see, e.g., Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009). To enhance the external validity of our findings, we partly adapted items from TIMSS 2011 (e.g., International Association for the Evaluation of Educational Achievement [IEA], 2013). The item adaption was necessary in order for all items to have the same formal structure: All items were presented in a multiple-choice format that comprised a short item stem, a separate one-sentence question, four answer options with one correct option (single-choice format) and, depending on the experimental condition, a picture. As well as the high relevance in various educational settings, another main reason for using multiple-choice items in this study was to prevent students' keyboard typing speed from influencing the TOT measures, inducing potential bias. Each text-only item was experimentally manipulated by adding a representational picture (RP) to the item stem, which illustrated important information given in the text, but never provided solution-relevant information beyond the text information (i.e., multiple representations; Ainsworth, 2006). All RPs were realistic schematic drawings in shades of gray, displayed under the verbal item stem (see Fig. 1). Both text-only and text-picture items were perfectly parallel apart from the added pictorial element. Three independent raters with a professional background in education concordantly confirmed this relation between the text and the picture for every item that was used in this study (i.e., 100% interrater reliability). All items confronted students with realistic situations, forcing them to apply their declarative science knowledge from biology, physics, and chemistry to everyday phenomena and problems. Thus, it was essential that the students correctly understood the situation in the item stem in order to be able to solve the problem correctly. The test was constructed to assess students' basic science achievement (scientific literacy). The



Text-Only Item

The pupils in our eyes change according to different external conditions. In Situation 1, Anna observes that Jana's pupils are very big. In Situation 2, Anna observes that Jana's pupils are much smaller.

What external condition has changed from Situation 1 to Situation 2?

Ⓐ The temperature in Situation 1 is higher.
Ⓑ The temperature in Situation 2 is higher
Ⓒ The light in Situation 2 is brighter.
Ⓓ The light in Situation 1 is brighter.

Text-Picture Item

The pupils in our eyes change according to different external conditions. In Situation 1, Anna observes that Jana's pupils are very big. In Situation 2, Anna observes that Jana's pupils are much smaller.

What external condition has changed from Situation 1 to Situation 2?

Ⓐ The temperature in Situation 1 is higher.
Ⓑ The temperature in Situation 2 is higher
Ⓒ The light in Situation 2 is brighter.
Ⓓ The light in Situation 1 is brighter.

Fig. 1. Item example displaying the experimental conditions (*text-only* vs. *text-picture*); material adapted from IEA (2013).

items had a mean word count of M = 74.9 words (SD_{words} = 24.2). The EAP/PV reliability of the test was estimated as 0.83 and was comparably high for both text-only (0.80) and text-picture items (0.81).

3.2.2. *Apparatus*

The items were presented on 28 identical Lenovo® laptops with 15-inch screens. Each student was given a laptop and a mouse. The software flexSURVEY 2.0 (Hartenstein, 2012) was used to present each item on a single screen; scrolling was never necessary. Providing an answer automatically forwarded the student to the next task. The laptops were prepared in the room before the students entered. Each laptop presented different booklets at random throughout the sessions and students were also randomly assigned to their working space. The given answers, response times (per item), and the item presentation sequence (i.e., item positions) were recorded in a separate logfile for each student.

3.3. *Procedure*

Two experienced test administrators conducted the study during lesson time. All sessions were attended by a teacher and lasted up to 90 minutes. Students were instructed with standardized written and oral advice on how to answer the science items, which included always providing an answer and weighted guessing in cases of doubt. An example item was solved collectively in the classroom. Students were informed that they would not have the possibility to return to an earlier question after choosing an answer. Thus, they were repeatedly encouraged to take all the time they needed to solve an item but to nevertheless work in a focused way throughout the test. This detailed introduction was given to ensure that the science test was perceived and worked on as a power test, in order to achieve unbiased variance in students' response behavior and to prevent speededness from influencing response times. There was no time constraint for completing the test. The study was part of a larger project, and thus, additional background measures (i.e., general cognitive and reading abilities, item ratings and a short student questionnaire) were assessed which are not reported in detail in the following, because they are not in the focus of the present research.

3.4. *Measures*

3.4.1. *Solution success*

We measured students' solution success (i.e., response correctness) per item to compare their success rates in the text-only items with the text-picture items, in order to investigate the replicability of the multimedia effect in testing in the present study.

3.4.2. *Time on task*

We measured students' TOT per item (in seconds) as an indicator of both the cognitive and motivational effects of RPs, depending on the analysis focus (i.e., considering students' test engagement and item positions; see Table 1). To clean the data and reduce the risk of extreme response times causing bias, we replaced TOT values that were higher than two standard deviations above the item mean with the value of two standard deviations above the item mean (e.g., Goldhammer et al., 2014). This procedure was performed with log-transformed time data at an item level and also separately for text-only and text-picture items. This prevented the trimming from inducing systematic changes in students' TOT with regard to potential outliers and the experimental manipulation. In this way, we replaced 0.3% of the TOT data. While outliers in response times *below* item means (reflecting RGB) are usually also considered to be a threat to validity (e.g., Finn, 2015; Kong et al., 2007; Wise, 2006; Wise & Kong, 2005), this was not the case in the present study, where RGB constituted a central outcome variable and covariate in the analyses. Accordingly, we did not replace values below item means.

3.4.3. *Rapid-guessing behavior*

For each trial, we identified whether students did or did not show RGB; this served as a negative indicator of students' current item-solving motivation and, thus, as an indicator of the motivational function of RPs. While various methods for identifying RGB have been discussed (e.g., Finn, 2015; Kong et al., 2007; Lee & Jia, 2014; Wise, 2006; Wise & Kong, 2005), several advantages led us to use the *normative threshold* (NT) method (Wise & Ma, 2012), which takes the item-specific mean into consideration when setting a threshold to identify RGB. This was important in order to reduce the risk of systematic bias in the RGB identification in text-only compared to text-picture items. Furthermore, the NT method allows different threshold percentages to be defined, which means that response times shorter than, for example, 10%, 15% or 20% of the average solution time of an item are classified as rapid guesses. When setting RGB thresholds, it is important to achieve a balance between identifying as many non-effortful responses as possible and avoiding the classification of effortful responses as RGB (e.g., Lee & Jia, 2014; Wise & Kong, 2005). We used three approaches to identify and cross-validate the optimal threshold in this study. We evaluated thresholds following the NT10, NT15 and NT20 criteria, but the NT20 method was instantly rejected because many thresholds turned out to be above 10 s, which is hard to label as *rapid* guessing (Wise & Ma, 2012). Comparing the distribution of log-transformed response times by visual inspection for each item, the NT15 threshold performed better in identifying RGB than the NT10, considering the inclusion of well-defined "RGB bumps" at the left end of the response time distribution (Kong et al., 2007; Lee & Jia, 2014; Wise & Kong,

Table 2
Descriptive means (*M*) and standard deviations (*SD*) for the three outcome parameters (1) solution success, (2) time on task (TOT), and (3) rapid-guessing behavior (RGB) regarding *text-picture* and *text-only* items; presented separately for groups of *engaged students* (PT-RGB = 0) and *less engaged students* (PT-RGB > 0) for different item positions (i.e., three blocks à 12 items) and across *all students*.

		PT-RGB[a] = 0 (Engaged Students)			PT-RGB[a] > 0 (Less Engaged Students)			All Students
Outcome Parameter	Item Type	Item position 1–12	Item position 13–24	Item position 25–36	Item position 1–12	Item position 13–24	Item position 25–36	Grand Mean
Solution Success	*Text-Only* M (*SD*)	0.55 (*0.50*)	0.57 (*0.50*)	0.58 (*0.50*)	0.40 (*0.49*)	0.38 (*0.49*)	0.30 (*0.46*)	0.49 (*0.50*)
	Text-Picture M (*SD*)	0.64 (*0.48*)	0.62 (*0.48*)	0.66 (*0.47*)	0.47 (*0.50*)	0.43 (*0.50*)	0.40 (*0.49*)	0.57 (*0.50*)
Time on Task	*Text-Only* M (*SD*)	43.96 (*25.94*)	41.05 (*21.71*)	40.53 (*21.21*)	37.53 (*27.05*)	29.27 (*25.67*)	18.46 (*20.02*)	36.63 (*25.07*)
	Text-Picture M (*SD*)	41.70 (*22.97*)	42.05 (*21.34*)	39.08 (*22.04*)	38.90 (*24.20*)	32.99 (*23.24*)	19.38 (*19.43*)	37.44 (*23.26*)
Rapid-Guessing Behavior	*Text-Only* M (*SD*)	–[b]	–[b]	–[b]	0.08 (0.28)	0.23 (*0.42*)	0.33 (*0.47*)	0.08[c] (*0.28*)
	Text-Picture M (*SD*)	–[b]	–[b]	–[b]	0.03 (0.17)	0.09 (*0.29*)	0.26 (*0.44*)	0.04 (*0.20*)

[a] PT-RGB = Person-total rapid-guessing behavior (i.e., PT-RGB > 0 means that students showed RGB at least once over the course of testing).
[b] By definition, no RGB trials were observed for engaged students (PT-RGB = 0).
[c] RGB values across *all students* reflect the total percentage of RGB trials in the text-only and text-picture items in this study.

2005). We also took item solution rates in potential RGB trials into account. Across all items, the NT10 method resulted in a mean correct response rate of 0.25, and the NT15 method in 0.26, which are both close to the expected a priori probability for random guesses among four options (0.25). Because it was the goal of the present study to investigate RGB in connection with other variables, we considered a slightly more liberal threshold to be preferable. Taking all the criteria into account, we applied the NT15 method,[3] resulting in an average item-specific threshold of $M = 5.6$ seconds ($SD = 1.4$). As response times were measured in seconds, we applied rounding rules to define the thresholds. Overall, 884 out of 14,303 trials (6.1%) were classified as RGB, ranging from 0% to 18% RGB per item.

3.4.4. Person-total RGB

We calculated the proportion of RGB trials at a student level[4] and used this person-specific RGB tendency as a negative indicator of students' test engagement, which served as a covariate in the solution-success and TOT analyses. We refer to this measure as *Person-Total RGB* (PT-RGB). Higher values reflect lower test engagement (i.e., more RGB; max = 1) while zero values (PT-RGB = 0) code the engaged reference group of students who never showed RGB.

3.5. Statistical analyses

We report descriptive data in Table 2 to give an impression of the manifest data structure in the groups of engaged students (PT-RGB = 0) and less engaged students (PT-RGB > 0) across item positions (1–36). However, these results must be interpreted with care because they do not control for the data structure that resulted from presenting the items in a multi-matrix design. For the inferential analyses, we tested all the effects of interest in one analysis model for each outcome variable by using (generalized) linear mixed-effects models (GLMMs/LMMs; for an introduction see, e.g., Snijders & Bosker, 2012). The models estimated the influence of the *multimedia manipulation* on students' (a) solution success, (b) TOT, and (c) RGB, while taking interactions with *students' RGB tendency* (PT-RGB) and with *item positions* as covariates into account[5] (see Table 3). A central advantage of the mixed-effects models is that they are able to handle the unbalanced cross-classified data structure from the rotated multi-matrix design we employed in our study. We took the data structure into account by defining fixed effects and random intercepts as well as random slopes regarding the picture manipulation for students and for items (i.e., allowing the multimedia effect to vary across both entities; see, e.g., Baayen, Davidson, & Bates, 2008). This flexibility of the models results in better estimates of standard errors and more valid statistical inferences (Barr, Levy, Scheepers, & Tily, 2013). In order to estimate students' solution success and RGB, we used GLMMs that can handle binary outcomes, similar to logistic regression analyses. All the fixed effects parameters γ can be interpreted similar to ordinary regression coefficients. We used the lme4 package (Bates, Maechler, & Bolker, 2012) in R (R Core Team, 2015) to conduct the mixed-effects analyses.

4. Results

4.1. Solution success

The descriptive data, presented in Table 2, show that students solved text-picture items more often than corresponding text-only items. Though the overall level of response correctness was higher for engaged students, a similar multimedia effect for both engaged and less engaged students was evident in the descriptive data; the differences in response correctness between text-only and text-picture items were highly comparable for items presented at the beginning, in the middle, and at the end of the test.

Testing our *multimedia effect in testing hypothesis (H1)*, as expected, a generalized mixed-effects model that predicted students' solution

[3] Conducting our analyses based on the NT10 criterion for setting RGB thresholds would not affect any of the reported results in this study.

[4] Among the 401 students, the response behavior of 256 students (63.8%) was classified as effortful solution behavior in all presented items (PT-RGB = 0). The remaining 138 students (36.2%) had heterogeneous scores, ranging from $\text{PT-RGB}_{\text{min}} = 0.03$ to $\text{PT-RGB}_{\text{max}} = 0.75$ ($M = 0.06$; $SD = 0.007$). A group of 65 students (16.2%) showed RGB in up to 10% of the items, 39 students (9.7%) showed RGB in more than 10% but less than 20% of the items, and 29 students (7.7%) ranged between 20% and 49% RGB throughout the test. Only 12 students (2.9%) skipped 50% or more items.

[5] Considering students' reading abilities as a covariate in the analyses showed that reading comprehension did not affect the main results and conclusions of our study in any outcome parameter (i.e., solution success, TOT, RGB). Thus, we refrain from reporting reading effects to keep the (G)LMM models concise.

Table 3
Model parameters that predict students' (a) *solution success*, (b) *time on task* (TOT), and (c) *rapid-guessing behavior* (RGB) in three (generalized) linear mixed-effects models.

		Solution Success				Time on Task				Rapid-Guessing Behavior			
Fixed effects		Estimate	*SE*	*z*	*p*	Estimate	*SE*	*t(44)*	*p*	Estimate	*SE*	*z*	*p*
	Intercept γ_0	0.21	0.13	1.61	.108	49.56	1.77	28.05	≤.001	−9.04	0.35	−25.93	≤.001
	Picture[a] γ_1	0.30	0.12	2.52	.012	−2.50	1.12	−2.23	.031	−0.89	0.24	−3.68	≤.001
	PT-RGB[b] γ_2	−2.15	0.41	5.21	≤.001	−26.46	5.57	−4.75	≤.001	–	–	–	–
	PT-RGB × Picture γ_3	−0.15	0.31	−0.49	.627	13.28	3.99	3.32	.002	–	–	–	–
	Position γ_4	−0.002	0.00	−0.39	.700	−0.48	0.05	−10.20	≤.001	0.13	0.01	14.04	≤.001
	Picture × Position γ_5	0.004	0.00	0.71	.479	0.12	0.05	2.54	.015	–	–	–	–
	PT-RGB × Position γ_6	−0.07	0.01	4.43	≤.001	−1.93	0.13	−15.12	≤.001	–	–	–	–
Random effects		*VAR*		*r*		*VAR*		*r*		*VAR*		*r*	
Students	Intercept	0.43				155.60				14.73			
	Picture Slope	–				70.24				3.26			
	Correlation			–				−0.44				−0.40	
Items	Intercept	0.25				66.46				0.34			
	Picture Slope	0.08				6.19				0.50			
	Correlation			0.24				0.19				−0.46	
	Residual	–				302.00				–			
	Deviance	17781.9				123980.5				3829.6			

Note. *SE* = Standard error; *VAR* = Variance; *r* = correlation; PT-RGB-related coefficients denote the expected variable change for a change from PT-RGB = 0 (0% RGB) to PT-RGB = 1 (100% RGB); these coefficients need to be broken down into smaller proportions for their interpretation (e.g., dividing the coefficient by five yields the expected change for a 0.2-unit increase in PT-RGB; i.e., for a student who engaged in RGB in 20% of the presented items, namely PT-RGB = 0.20).

[a] Dummy coding of the experimental conditions (0 = *text-only*; 1 = *text-picture*).

[b] Higher PT-RGB (person-total rapid-guessing behavior) values indicate a higher tendency to engage in RGB (PT-RGB = 0 means no RGB observed).

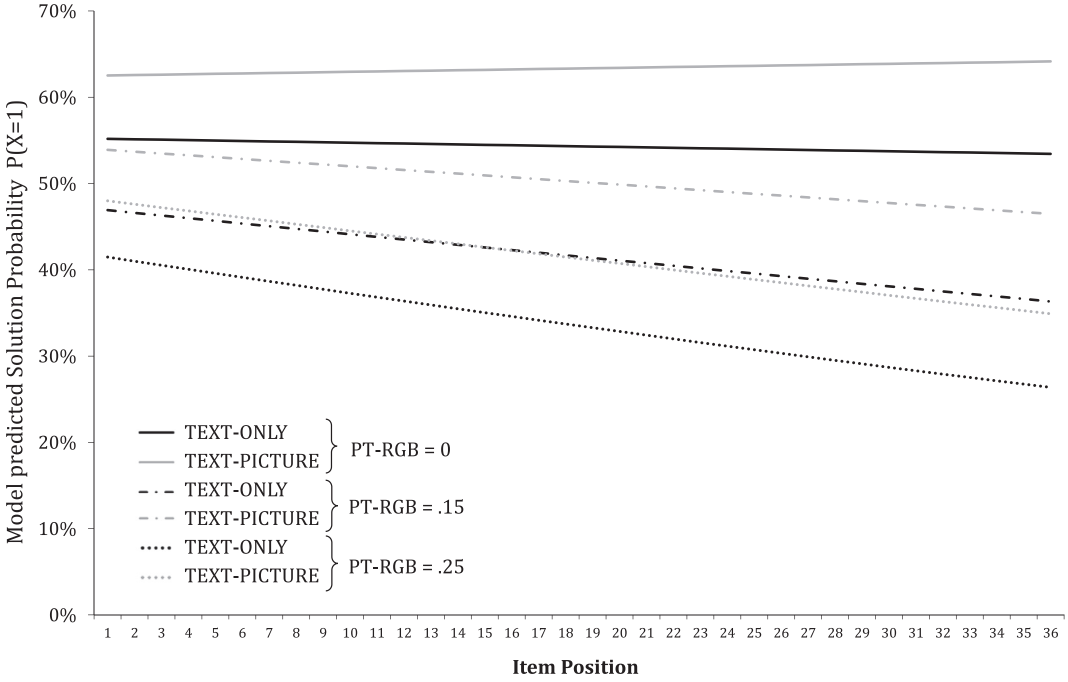


Fig. 2. Model-estimated solution success (probability) for students with different manifestations of PT-RGB (person-total rapid-guessing behavior) for *text-only* (black lines) and *text-picture* (gray lines) items across testing time (i.e., item positions).

success (see Table 3) revealed a significant positive main effect for the factor *picture* (γ_1 = 0.30; z = 2.52; p = .012), reflecting a multimedia effect in testing and supporting *Hypothesis H1*. Addressing our explorative questions, we found a significant main effect for the factor *PT-RGB* (γ_2 = −2.15; z = 5.21; $p \leq$ .001), reflecting a negative impact of students' RGB tendency on their solution success, while the interaction effect of *PT-RGB* × *picture* was not significant (γ_3 = −0.15; z = −0.49*;* p = .626). Considering potential item position effects, we found both the effect of the factor *item position* (γ_4 = −0.002; z = −0.39; p = .700) and the interaction effect of *item position* × *picture* (γ_5 = 0.004; z = 0.71; p = .479) to be nonsignificant. However, a significant negative interaction effect of *PT-RGB* × *item position* (γ_6 = −0.07; z = 4.43; $p \leq$ .001) occurred, suggesting that the solution success of less engaged students (i.e., PT-RGB > 0) decreased with increasing item positions (i.e., *item position effect*). The three-way interaction of *PT-RGB* × *picture* × *item position* was nonsignificant

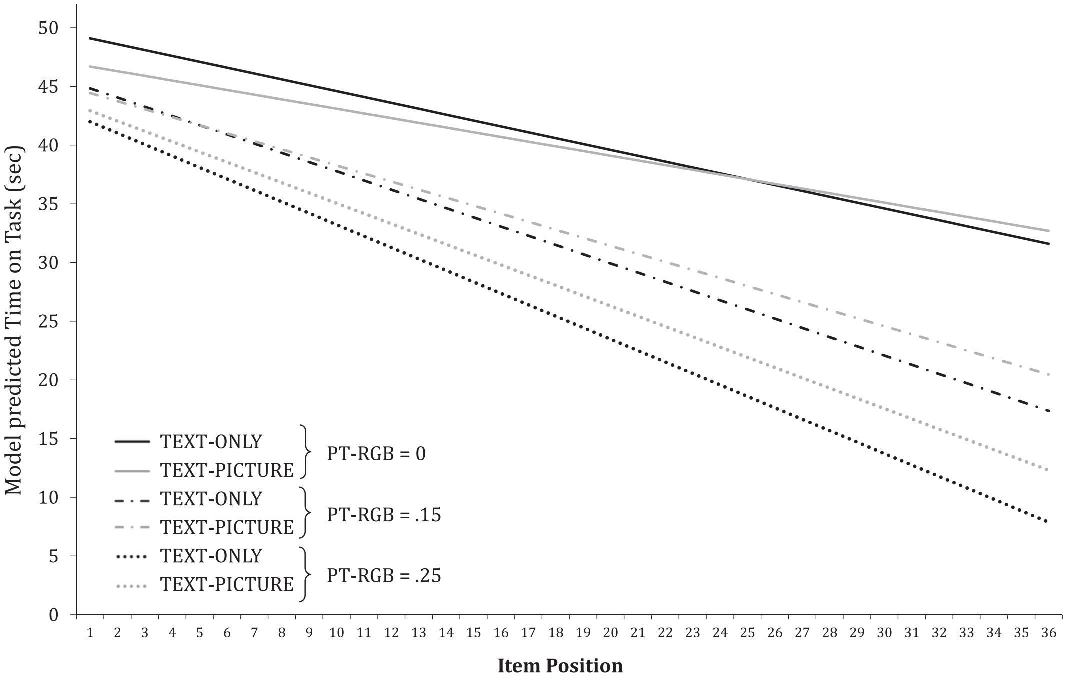


Fig. 3. Model-estimated time on task (TOT) in seconds for students with different manifestations of PT-RGB (person-total rapid-guessing behavior) for text-only (black lines) and text-picture (gray lines) items across testing time (i.e., item positions).

(p = .482) and, thus, not integrated into the final model. Finally, it is notable that the variance of the random effect of the picture slope for students was very small (with a slope-intercept correlation of $r = 1$), suggesting that the multimedia effect did not vary much across students. Therefore, we did not include the random effect of the picture slope into the final model. All model parameters are displayed in Table 3; additionally, Fig. 2 visualizes the model-predicted effects on students' solution success across item positions for selected examples of students' with different PT-RGB values (0, 0.15 and 0.25).

4.2. Time on task

The descriptive data in Table 2 show that as expected, engaged students on average needed less TOT to solve text-picture compared to text-only items in the first test block. However, across all trials and students (grand mean), the TOT spent on text-picture items was slightly higher than the TOT spent on text-only items.

Testing our *cognitive facilitation hypothesis (H2)* in a linear mixed-effects model (see Table 3) revealed a significant negative main effect of the factor *picture* ($\gamma_1 = -2.50$; $t = -2.23$; $p = .031$) on students' TOT. Following the regression coefficient interpretation rules[6] and the parameters we included in the model (see Table 3), this effect predicted a shorter TOT in text-picture items compared to text-only items for the group of engaged students (i.e., PT-RGB = 0) at the test start (i.e., item position = 0), supporting our *Hypothesis H2*. Furthermore, our explorative analyses showed a significant negative main effect of *PT-RGB* ($\gamma_2 = -26.46$; $t = -4.75$; $p \leq .001$), indicating that students who engaged in RGB spent less TOT on solving the test items; this was to be anticipated, given the relative interdependence of PT-RGB and TOT. However, a positive interaction effect of *PT-RGB* × *picture* ($\gamma_3 = 13.28$; $t = 3.32$; $p = .002$) revealed that less engaged students spent more TOT when working on text-picture items compared to text-only items. We also found a significant main effect of *item position* ($\gamma_4 = -0.48$; $t = -10.20$; $p < .001$), which predicted a decrease in TOT over the course of testing even for engaged students, while a significant interaction effect of *PT-RGB* × *item position* ($\gamma_6 = -1.93$; $t = -15.12$; $p \leq .001$) revealed that the TOT decrease over the course of the test was much stronger for less engaged students compared to engaged students. A significant positive interaction effect of *picture* × *item position* ($\gamma_5 = 0.12$; $t = 2.54$; $p = .015$) indicated that RPs reduced the TOT decrease that occurred with increasing item positions (see γ_4). The three-way interaction, *PT-RGB* × *picture* × *item position*, was nonsignificant (p = .281) and, thus, not integrated into the final model. All model parameters are displayed in Table 3; Fig. 3 provides a visualization of the predicted effects for selected examples of students' PT-RGB values (0, 0.15, and 0.25) to visualize the model-implied interplay between all factors that significantly influenced students' TOT.

4.3. Rapid-guessing behavior

Our descriptive data in Table 2 show that RPs reduced students' tendency to engage in RGB to a substantial extent. RGB rates increased with item positions, but text-picture items were on average always less affected by RGB than text-only items.

Supporting our *motivational function hypothesis H3a,* a significant negative main effect for the factor *picture* ($\gamma_1 = -0.89$; $z = -3.68$; $p \leq .001$) occurred in the generalized mixed-effects model (see Table 3), reflecting an RGB-reducing influence of RPs, as expected. Furthermore, the significant main effect of the factor *item position* ($\gamma_2 = 0.13$; $z = 14.04$; $p \leq .001$) revealed that students' RGB rates increased with each item position over the course of the testing time. However, rejecting our *motivational function hypothesis H3b*, the anticipated interaction effect *picture* × *item position* was nonsignificant (p = .879) and, thus, not part of the final model.

5. Discussion

The functions of pictorial elements in testing material have just recently started to receive attention in educational research. While first studies repeatedly demonstrated that RPs have a positive impact on students' performance in assessment, the underlying mechanisms of this

[6] The standard interpretation of effects in a regression model with interaction effects pertains to the hypothetical situation, in which all other variables are held constant at zero.

multimedia effect in testing still need more investigation. In the light of assumptions from cognitive multimedia and interest theories, we proposed that RPs have a *cognitive* as well as a *motivational* function and that these functions work together to foster students' performance. In the following, we discuss our findings in relation to our research hypotheses. After that, we integrate the results into an overall conclusion and suggest some educational implications. Finally, we reflect on the study's limitations and propose topics for future research.

5.1. The multimedia effect in testing

Confirming our *multimedia effect in testing hypothesis (H1)*, our results replicated earlier findings that students are more successful in solving text-picture compared to corresponding text-only items (e.g., Lindner et al., 2016; Saß et al., 2012). Furthermore, the explorative analyses showed that less engaged students were overall less successful in solving the test items compared to engaged students. However, there was no significant interaction between RPs' presence and students' test engagement (i.e., PT-RGB), indicating that the multimedia effect in testing did not change to a meaningful extent with different levels of test engagement. This result was supported by a very small and non-significant variance component of the random slope (picture) effect for students, which indicated that there was no substantial variation in the multimedia effect at a student level. Overall, these findings match the results of Lindner et al. (2016), showing that RPs' positive impact on performance hardly varied across students. Nevertheless, such a finding, with a focus on students with a lower test engagement, is interesting and was not necessarily expected.

Exploring students' solution success across the test, a significant performance decrease over the course of testing (i.e., item position effect; e.g., Hartig & Buchholz, 2012) was only evident for less engaged students, but not for engaged students. Most importantly, however, we did not find evidence that RPs significantly influenced students' performance in a different manner across item positions. This applied to both engaged and less engaged students. Thus, taken together, a substantial multimedia effect was present for all students and it neither changed to a significant extent as a function of students' level of test engagement nor as a function of item positions (see Fig. 2). The multimedia effect in testing thus emerges as a stable and predictable effect, not only across studies (Hartmann, 2012; Lindner et al., 2016, 2017; Saß et al., 2012) but also across students, items, and testing time.

5.2. A cognitive facilitation function

To test our *cognitive facilitation hypothesis (H2)* concerning RPs' impact on item processing, we focused on the TOT engaged students spent on solving text-only compared to text-picture items at the test start. As explained in the Theory section, the reason for this was that we expect only the TOT of motivationally fully engaged students to provide a valid reflection of the actual processing time necessary to solve the items and, thus, to allow tracing cognitive facilitations in their TOT. Both the descriptive data and the mixed-effects model parameters show that as expected, engaged students solved text-picture items significantly faster (about 2.5 seconds) compared to text-only items at the test start, confirming hypothesis H2. More specifically, the mixed-effects model predicts that the acceleration effect would be traceable until the item position 23 (see also Fig. 3), but this prediction should not be interpreted in a definite manner. Overall, even though the effect decreased with increasing item positions (probably reflecting a motivational effect of RPs; see discussion below), the data support the assumption that RPs have the potential to accelerate important processes in solving text-picture items, even though the maximum potential becomes only evident when students' test engagement and item position effects on TOT are controlled for. This finding suggests that RPs' beneficial cognitive effects can compensate for even more than just the additional encoding time that is necessary to process the added RP.

Building on the assumptions of the CTML and the ITPC in the learning context (e.g., Eitel et al., 2013; Mayer, 2005; Schnotz & Bannert, 2003), together with the eye-tracking results obtained by Lindner et al. (2017) in testing, such acceleration probably relates to a faster processing of the text item-stem information (i.e., easier mental model construction) and an easier updating of the mental model by using the RP when considering solutions for the presented problem (Schnotz et al., 2014). As the present results are in line with previous eye-tracking data, it is likely that the cognitive facilitations described by Lindner et al. (2017) not only occur under laboratory conditions but can also be transferred to an everyday classroom setting. Combined with the successful replication of the multimedia effect in testing, our data corroborate the proposition that RPs play a substantial role in supporting students in processing the item information more efficiently and in better applying their conceptual knowledge to a presented problem, for example, by preventing incorrect conclusions based on misinterpretations of the item-stem content. In accordance with this, RPs may have also reduced extraneous cognitive load. Such cognitive facilitations of RPs could be especially important in tests that present complex information, but they were apparently also helpful with regard to the basic scientific concepts that were assessed by the test in our study.

5.3. A motivational function

Testing our *motivational function hypothesis (H3a)*, we found that text-picture items were significantly less prone to RGB compared to corresponding text-only items, which we interpret in terms of an increased motivational effort. Confirming hypothesis H3a, our data indicate that RPs attracted students to become engaged with a task rather than to skip it without investing any effort, which is in line with the explorative findings of Wise et al. (2009). This may also indicate that RPs fostered students' situational interest and their test-taking motivation, which matches findings that students report more pleasure when solving text-picture items compared to text-only items (Lindner et al., 2016). However, both text-only and text-picture items were prone to increasing rates of RGB over the course of the test (i.e., item positions), which suggests that students' motivation to engage in effortful solution behavior decreased over testing time. This interpretation was also supported by our explorative part of the TOT analysis, which revealed that all students, but especially less engaged students, showed a decrease in TOT with increasing item positions; but this decrease was significantly smaller for text-picture items compared to text-only items (cf. Fig. 3). Thus, RPs seemed to have counteracted a decrease in motivational effort in the course of testing for all students.

However, rejecting our *motivational function hypothesis (H3b),* in contrast to the results obtained by Wise et al. (2009), we did not find a significant interaction between the RGB-reducing effect of RPs and increasing item positions. Hence, text-picture items were generally less affected by RGB, but not to a higher extent in items presented in later positions. Taking into consideration the fact that the interaction effect demonstrated by Wise et al. (2009) was found in a test consisting of 63 items, the 36 items in our study might not have been enough to show the effect. However, our descriptive data (see Table 2) indicate that RPs increasingly helped to prevent RGB within a certain time frame (i.e., in the first two test blocks), while RPs' efficacy did not seem to increase linearly in later item positions (i.e., we found a smaller RGB-reducing effect of RPs in the last test block). This might indicate that RPs attracted students' attention for a certain amount of testing time, but lost their attraction when students got used to solving text-picture items. As the test of Wise et al. (2009) only included a few graphics, students probably did not get used to the graphics, and this might have supported the finding of a significant interaction effect of *picture* × *item position* on students' tendency to show RGB. Overall, the difference between our results and those of Wise et al. (2009) indicates that future research needs to focus more explicitly on such framing factors.

5.4. Conclusions and educational implications

This study provides behavior-related empirical insights into the composition of the multimedia effect in testing, which it was able to replicate. Reviewing our findings together with earlier research, there is clear evidence that RPs have beneficial (1) *cognitive* effects on students' item processing, as well as beneficial (2) *motivational* effects, reflected in an RP-related improvement in less engaged students' test-taking behavior (i.e., reduced RGB and increased TOT). However, RPs did not reduce RGB to a higher extent in later item positions, which is not in line with the results reported by Wise et al. (2009) and thus needs to be clarified in future studies.

Our findings are not only relevant for educational researchers, but also for practitioners, such as test constructors in LSA studies, or even for teachers. This is because, in line with earlier studies, our data suggest that RPs could be deliberately applied in testing material to achieve certain goals, such as facilitating the cognitive processing of the item information in order to prevent students from misinterpreting the testing material, or to help unmotivated students to overcome their tendency to engage in RGB and to increase the time they spend working on the test items. All of these aspects can be considered as potential factors in promoting more reliable responses and reducing construct-irrelevant variance (see, e.g., Haladyna & Downing, 2004) and could, thus, also promote a more valid interpretation of test scores. Accordingly, integrating RPs into test items may serve as a design principle for the purpose of improving item attractiveness and triggering students' situational interest and motivation. Enhancing students' work ethic in such a way could be particularly beneficial in low-stakes assessments, where motivational issues and the need to trigger and maintain students' willingness to work on a test play a crucial role in obtaining reliable test results (Baumert & Demmrich, 2001; Wise & DeMars, 2005).

However, RPs' potential to help students encode the item-stem information and to build a mental model more easily may, however, not always be preferable and may not necessarily contribute to obtaining more appropriate test scores. This is because, in contrast to learning situations, where fostering students' understanding of the material is always a central goal, building a mental model based on a text (i.e., to interpret the text correctly) may be an important facet of the construct measured in testing. Thus, taking away the need to build a mental model from scratch by adding an RP to a test item might remove that facet from the test and could thereby undermine a test's construct validity (e.g., content-, cognitive- or structural validity; see, e.g., Messick, 1995). However, this issue cannot be addressed at a global level and future research is clearly needed to further investigate the relation of RPs and test validity. Until then, this aspect must be reviewed individually by test constructors with regard to the construct that they intend to measure. Accordingly, even though RPs' motivational function most probably contributes to obtaining more reliable test results, independent of the measured construct, it would not be wise to make a general recommendation regarding the use of RPs in test items.

Furthermore, the context of using RPs in testing also needs to be considered, as the resources for construction and editing of RPs are, for example, much lower in teacher-constructed classroom tests compared to standardized LSA tests. This might affect the quality of RPs and, thus, their effectiveness. Also the answer format of the test items should be considered when using RPs, as there is no empirical evidence yet that RPs have the same effects in constructed response items as shown for multiple-choice items. Nonetheless, keeping an eye on these issues, test constructors in LSA studies and teachers should feel encouraged to consider using RPs more often and in a more targeted manner.

5.5. Limitations and future directions

First, it is to note that the interrelation between the outcome parameter TOT and the covariate PT-RGB both refer in a different way to the time that students spent on solving the test items. We used the PT-RGB measure to identify engaged and less engaged students because RGB is a reliable indicator of students' test engagement and is free of social desirability (Finn, 2015; Wise & Kong, 2005). Nevertheless, due to the way in which the measures are connected to each other, the PT-RGB *main effect* in the TOT analyses needs to be interpreted with some care. In particular, the observation of lower TOT levels for less engaged students reflected the operationalization to some extent, as students who show more RGB should, by definition, also have a shorter TOT on average. However, this only affects the interpretation of the PT-RGB *main effect*, but not that of the *interaction effects* regarding the picture manipulation or the item positions, which were in the focus of interest. Finally, it should be noted that, as we did not manipulate students' engagement levels experimentally, our explorative effects regarding this factor can only be interpreted in a correlative manner.

Second, our sample was not necessarily representative for students in general, as we tested fifth and sixth grade students who might have, for example, especially enjoyed the presence of RPs. However, the students might also have had problems with integrating the text and picture information properly, which is typical for younger students (see, e.g., Ainsworth, 2006) and could potentially influence the results. Thus, the generalizability of our findings should be reconsidered, testing older students.

Third, the items in this study cannot be considered to be representative of test items in general, because we used relatively easy multiple-choice science items that required controlled cognitive processing (cf. Goldhammer et al., 2014) and had a rather low amount of text. This may, of course, limit the generalization of the findings to other tasks. In particular, the extent to which multimedia effects in testing transfer across content domains, task requirements or item formats (e.g., open response) is not yet clear. This is because—as in our study—all studies conducted in this area so far have used multiple-choice science items with RPs. Thus, the results are certainly reliable for this type of science tasks, but items with longer texts, in other content domains, with different RPs (e.g., dynamic or colored versions) and with different item formats should be systematically studied in future research. Furthermore, the results can also not be easily transferred to other types of visualizations, such as decorative pictures, tables, graphs, or diagrams (Ainsworth, 2006; Carney & Levin, 2002; Mayer, 2005). Hence, future research should also focus on different visualizations in order to gain a sound understanding of psychometric, cognitive, and motivational multimedia effects in testing to make these effects more predictable for item constructors.

Acknowledgements

We thank Gráinne Newcombe for language editing, Benjamin Strobel for programming support and together with Julia Barenthien for helping with data collection.

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*, 163–181. http://dx.doi.org/10.1037/a0015719.

Ackerman, P. L., Kanfer, R., Shapiro, S. W., Newton, S., & Beier, M. E. (2010). Cognitive fatigue during testing: An examination of trait, time-on-task, and strategy influences. *Human Performance, 23*(5), 381–402. http://dx.doi.org/10.1080/08959285.2010.517720.

Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94*, 545–561. http://dx.doi.org/10.1037/0022-0663.94.3.545.

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. http://dx.doi.org/10.1016/j.jml.2007.12.005.

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*, 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using Eigen* *and S4 (R Package Version 0.999999–0)* [Computer software]. Retrieved from < https://CRAN.R-project.org/package=lme4 > .

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441–462. http://dx.doi.org/10.1007/BF03173192.

Bennett, R. E., Goodman, M., Hessinger, J., Kahn, H., Ligget, J., Marshall, G., & Zack, J. (1999). Using multimedia in large-scale computer-based testing programs. *Computers in Human Behavior, 15*, 283–294. http://dx.doi.org/10.1016/S0747-5632(99)00024-2.

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98*, 182–197. http://dx.doi.org/10.1037/0022-0663.98.1.182.

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

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA Reading Assessment. *Journal of Educational and Behavioral Statistics, 39*, 502–523. http://dx.doi.org/10.3102/1076998614558485.

Eitel, A., & Scheiter, K. (2015). Picture or text first? Explaining sequence effects when learning with pictures and text. *Educational Psychology Review, 27*, 153–180. http://dx.doi.org/10.1007/s10648-014-9264-4.

Eitel, A., Scheiter, K., Schüler, A., Nyström, M., & Holmqvist, K. (2013). How a picture facilitates the process of learning from text: Evidence for scaffolding. *Learning and Instruction, 28*, 48–63. http://dx.doi.org/10.1016/j.learninstruc.2013.05.002.

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series, 1–17*. http://dx.doi.org/10.1002/ets2.12067.

Glenberg, A. M., & Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language, 31*, 129–151. http://dx.doi.org/10.1016/0749-596X(92)90008-L.

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

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

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27. http://dx.doi.org/10.1111/j.1745-3992.2004.tb00149.x.

Hartenstein, S. (2012). *flexSURVEY – Flexible PHP-driven online surveys (Version 2.0)* [Computer software]. Retrieved from < http://www.flexsurvey.de/Download > .

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54*, 418–431. Retrieved from <http://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2012_20121224/06_Hartig.pdf>.

Hartmann, S. (2012). *Die Rolle von Leseverständnis und Lesegeschwindigkeit beim Zustandekommen der Leistungen in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenz*. [*The role of reading comprehension and reading speed in text-based assessments of scientific inquiry skills*]. Doctoral dissertation. University of Duisburg-Essen. Retrieved from < http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-33260/hartmann_diss.pdf > .

Hidi, S. (2006). Interest: A unique motivational variable. *Educational Research Review, 1*, 69–82. http://dx.doi.org/10.1016/j.edurev.2006.09.001.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111–127. http://dx.doi.org/10.1207/s15326985ep4102_4.

International Association for the Evaluation of Educational Achievement [IEA] (2013). *TIMSS 2011 assessment released science items*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College/International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat. Retrieved from < http://nces.ed.gov/timss/pdf/TIMSS2011_G4_Science.pdf > .

Inzlicht, M., & Schmeichel, B. J. (2012). What is ego depletion? Toward a mechanistic revision of the Resource Model of Self-Control. *Perspectives on Psychological Science, 7*, 450–463. http://dx.doi.org/10.1177/1745691612454134.

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

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*, 606–619. http://dx.doi.org/10.1177/0013164406294779.

Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education, 2*(8), 1–24. http://dx.doi.org/10.1186/s40536-014-0008-1.

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

Leutner, D. (2014). Motivation and emotion as mediators in multimedia learning. *Learning and Instruction, 29*, 174–175. http://dx.doi.org/10.1016/j.learninstruc.2013.05.004.

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*, 92–102. http://dx.doi.org/10.1016/j.learninstruc.2016.10.007.

Lindner, M. A., Ihme, J. M., Saß, S., & Köller, O. (2016). How representational pictures enhance students' performance and test-taking pleasure in low-stakes assessment. *European Journal of Psychological Assessment,* 1–10. http://dx.doi.org/10.1027/1015-5759/a000351 Advance online publication.

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

Mayer, R. E. (Ed.) (2005). *The Cambridge handbook of multimedia learning*. Cambridge, England: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511816819.005.

Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learning and Instruction, 29*, 171–173. http://dx.doi.org/10.1016/j.learninstruc.2013.04.003.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741.

Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology, 85*, 424–436. http://dx.doi.org/10.1037/0022-0663.85.3.424.

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review, 19*, 309–326. http://dx.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). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College. Retrieved from < http://files.eric.ed.gov/fulltext/ED512411.pdf > .

Penk, C., & Richter, D. (2016). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability,* 1–25. http://dx.doi.org/10.1007/s11092-016-9248-7 Advance online publication.

R Core Team (2015). R: A language and environment for statistical computing and graphics. [Computer software]. Retrieved from <http://www.R-project.org>.

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

Schiefele, U. (2009). Situational and individual interest. In K. Wentzel, A. Wigfield, & D. Miele (Eds.). *Handbook of motivation at school* (pp. 197–222). New York, NY: Routledge.

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

Schnotz, W., & Kürschner, C. (2008). External and internal representations in the acquisition and use of knowledge: Visualization effects on mental model construction. *Instructional Science, 36*, 175–190. http://dx.doi.org/10.1007/s11251-007- 9029-2.

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

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190. http://dx.doi.org/10.1037/0033-295X.84.2.127.

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

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296. http://dx.doi.org/10.1023/A:1022193728205.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Vekiri, I. (2002). What is the value of graphical displays in learning? *Educational Psychology Review, 14*, 261–312. http://dx.doi.org/10.1023/A:1016064429161.

Wirth, J. (2008). Computer-based tests: Alternatives for test and item design. In J. Hartig, E. Klieme, & D. Leutner (Eds.). *Assessment of competencies in educational contexts* (pp. 235–252). Göttingen: Hogrefe.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*, 95–114. http://dx.doi.org/10.1207/s15324818ame1902_2.

Wise, S. L., & Ma, L. (2012, April). 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*.

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

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*, 27–41. http://dx.doi.org/10.1080/10627191003673216.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163–183. http://dx.doi.org/10.1207/s15324818ame1802_2.

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

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

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337–362. http://dx.doi.org/10.1207/S15324818AME1504_02.