



How Representational Pictures Enhance Students' Performance and Test-Taking Pleasure in Low-Stakes Assessment

Marlit A. Lindner, Jan M. Ihme, Steffani Saß, and Olaf Köller

Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Abstract: Pictures are often used in standardized educational large-scale assessment (LSA), but their impact on test parameters has received little attention up until now. Even less is known about pictures' affective effects on students in testing (i.e., test-taking pleasure and motivation). However, such knowledge is crucial for a focused application of multiple representations in LSA. Therefore, this study investigated how adding *representational pictures* (RPs) to text-based item stems affects (1) item difficulty and (2) students' test-taking pleasure. An experimental study with $N = 305$ schoolchildren was conducted, using 48 manipulated parallel science items (text-only vs. text-picture) in a rotated multimatrix design to realize within-subject measures. Students' general cognitive abilities, reading abilities, and background variables were assessed to consider potential interactions between RPs' effects and students' performance. Students also rated their item-solving pleasure for each item. Results from item-response theory (IRT) model comparisons showed that RPs only reduced item difficulty when pictures visualized information *mandatory* for solving the task, while RPs substantially enhanced students' test-taking pleasure even when they visualized *optional* context information. Overall, our findings suggest that RPs have a positive cognitive and affective influence on students' performance in LSA (i.e., *multimedia effect in testing*) and should be considered more frequently.

Keywords: multimedia effect, multiple representations, low-stakes assessment, item-writing guidelines, test-taking motivation

The political power of standardized educational assessment places enormous demands on a valid test construction and interpretation. Accordingly, item-writing principles play a major role in the optimization of tests (Haladyna, Downing, & Rodriguez, 2002; Haladyna & Rodriguez, 2013). However, so far, little attention has been devoted to the integration of pictorial elements into test items. This is somewhat unexpected, because pictures constitute eye-catching elements in many instruments measuring education-related competencies in large-scale assessment (LSA) studies, such as TIMSS (*Trends in International Mathematics and Science*; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009) or PISA (*Programme for International Student Assessment*; OECD, 2013). Studying the impact of different types of visualizations on item difficulty is essential if pictorial elements are to be used in a more targeted manner and based on empirical findings rather than on the individual theories of test constructors. Moreover, pictures might also impact students' affective state (e.g., Lenzner, Schnotz, & Müller, 2013) when solving test items, thereby constituting a potential variable for improving test validity in low-stakes testing by enhancing students' test-taking pleasure and motivation. In view of this, the present study investigated the effects of *representational pictures*

(RPs) on (1) *item difficulty* and on (2) *students' test-taking pleasure*, because RPs are frequently used in the context of science assessment, for example, to clarify item stem information or to put the task in a realistic context (cf. Mullis et al., 2009; OECD, 2009).

Representational pictures (RPs) visualize information from a corresponding text (Vekiri, 2002). Thus, together with the verbal information (written or spoken), they constitute *multiple representations* (Ainsworth, 1999) by providing two independent representational codes for mental model construction (*Cognitive Theory of Multimedia Learning* [CTML], Mayer, 2005; *Integrated Model of Text and Picture Comprehension* [ITPC], Schnotz & Bannert, 2003). Building a coherent situational mental model equals understanding external learning (or testing) material correctly (cf. Eitel, Scheiter, Schüler, Nyström, & Holmqvist, 2013; Schnotz & Bannert, 2003), thereby establishing the base for learning and for problem-solving in testing situations. A high number of studies provide evidence for a beneficial effect of RPs on students' performance when *learning* with text and pictures (*multimedia learning*; see, e.g., Levie & Lentz, 1982; Vekiri, 2002). These effects are explained by cognitive multimedia theories (CTML, ITPC) that, in a nutshell, refer to a dual coding of information in separate

verbal and visual processing channels in working memory (Baddeley, 1986; Paivio, 1986) and to a positive impact on the speed (Eitel et al., 2013) and accuracy of mental model construction (Schnotz & Bannert, 2003; Schnotz et al., 2014). In contrast, research focusing on how RPs affect processes and outcomes in testing situations is only at its very beginning.

In the following, we refer to RPs' influence on test outcome parameters as *Multimedia Effect* (MME) in testing. This term is usually used in the instructional sciences to describe the beneficial influence of pictures on learning performance (cf. Mayer, 2005). However, the term is just as convenient in multimedia testing, as recent studies have provided evidence that RPs can also have a positive influence on students' performance when integrated into test items (Hartmann, 2012; Prenzel, Häußler, Rost, & Senkbeil, 2002; Saß, Wittwer, Senkbeil, & Köller, 2012); this was reflected in higher test scores and a corresponding decrease in item difficulty. Accordingly, we propose that assumptions from multimedia theories (Mayer, 2005; Schnotz & Bannert, 2003) can be cautiously transferred to testing situations, in which RPs may likewise support efficient information processing and mental model construction to help students understand and solve the presented problem correctly. However, the relevance of RPs for solving the task may also influence the extent to which they affect students' performance. External representations should be especially helpful when they allow students to build a task-relevant mental model that can be directly applied for necessary mental transformations or for drawing conclusions to solve an item. In contrast, optional information would not help students to build a task-relevant mental model, but could nevertheless reactivate their background knowledge. In the worst case, pictures representing optional information might deflect students from concentrating on the main task. Accordingly, one goal of the present study was to investigate MME effect sizes for RPs that visualize information which is mandatory for correctly solving the presented problem, and for RPs that visualize optional context information.

Furthermore, little is known about interactions between students' characteristics and the influence of RPs on students' performance in testing (i.e., MME). Learning more about this might allow test constructors to integrate RPs in a more goal-oriented manner into assessment, for example, to support students with specific needs. This is because RPs can serve as an alternative, more easily accessible source of information for understanding a problem and building a coherent situational mental model (Eitel et al., 2013; Schnotz & Bannert, 2003; Schnotz et al., 2014). This might especially help students with poor reading abilities to demonstrate their science competence. A study by Hartmann (2012) provides tentative evidence

for this assumption: it showed that replacing item stem text blocks by RPs was especially beneficial for slow readers. Reproducing this effect is desirable because reducing reading demands, for example, in a science test, means reducing construct-irrelevant variance in favor of construct validity (cf. Messick, 1989). Furthermore, there is evidence from multimedia learning that students' abilities can interact with MMEs (e.g., Levie & Lentz, 1982; Schnotz et al., 2014), and this may also apply to multimedia testing. Thus, to gain first insights into student characteristics that may moderate MMEs, we considered students' *background variables* (e.g., age, school track), *abilities* (e.g., general cognitive abilities, reading ability, grades), and *motivational attributes* (e.g., test-taking pleasure, test-effort, domain interest) in order to identify possible individual differences in how students' performance benefits from RPs.

Moreover, there is a research gap concerning the effects of RPs on students' test-taking pleasure and motivation, while a positive influence is often implicitly assumed when integrating pictures into low-stakes assessment. First evidence in favor of this assumption can be found in a study by Wise, Pastor, and Kong (2009), showing that (undefined) graphical elements were associated with reduced rapid-guessing behavior. This means that students decided to skip items without putting any effort into solving the task less often when a graphical element was present. Hence, pictorial elements seem to have at least short-term positive influences on strongly unmotivated students, while this does not necessarily mean that they also have a positive influence on averagely motivated students. Thus, the main goal of the present study was to systematically assess how RPs affect students' affective state and their test-taking pleasure as a proxy for their motivation to engage in solving the items. A finding that proves that RPs have a positive influence would be good news, especially for low-stakes LSA, which frequently faces problems with students' test-taking motivation (cf. e.g., Wise & DeMars, 2005).

To address the outlined gaps in research, we implemented an experimental study, using manipulated multiple-choice (MC) science items in an item-response theory (IRT) framework. With this, we wanted to address the following research questions:

Research Question 1 (RQ1): Is the MME (RPs reduce item difficulty) replicable in the present study? Does the effect depend on item characteristics and/or on the relevance of the information visualized by the RPs?

Research Question 2 (RQ2): Do students' characteristics interact with the MME?

Research Question 3 (RQ3): Do RPs enhance students' test-taking pleasure?

Methods

Sample and Study Design

The sample comprised $N = 305$ 5th- and 6th-grade students (47% female, $M_{\text{age}} = 11.55$, $SD_{\text{age}} = 0.77$) from three schools in the northern part of Germany. A total of $n = 221$ students attended academic track schools (Gymnasium) and $n = 84$ students attended a nonacademic track school (regional school). The test was a paper-pencil test, presented in eight different booklets that followed a rotated multimatrix design for a set of 96 experimentally manipulated science items (for design details, see Electronic Supplementary Material ESM 1). For this, we manipulated 48 basic text items (text-only) by adding a representational picture to the item stem (text-picture). Apart from that, items were perfectly parallel. The items were grouped into four blocks of 12, and each booklet contained three blocks (36 items) of which at least one block (12 items) was presented as a text-only and one as a text-picture version, so that each student answered items under both experimental conditions while never answering the same item twice (within-subject design). The multimatrix design perfectly balanced the presentation order of the different blocks of text-only and text-picture items to prevent bias being caused by item position effects. At one regional school, booklets were shortened to 24 items (two blocks) because of a time restriction set by the school, which was unproblematic in our study design (cf. ESM 1). Booklets were randomly assigned to the students.

Measures

Further details concerning the instruments (e.g., items, descriptive statistics) are provided in the Electronic Supplementary Material 2 (ESM 2).

Background Questionnaire

Students' demographics, their last grade in German and science, their science interest (five items, $\alpha = .83$; adapted from PISA 2006/TIMSS 2011; cf. OECD, 2009; Martin & Mullis, 2012), and their reading self-concept (four items, $\alpha = .65$; Möller & Bonerad, 2007) were assessed (interest and self-concept on a 4-point Likert scale; *strongly agree* to *strongly disagree*).

Manipulated Science Items

Forty-eight multiple-choice items were constructed in close connection to the TIMSS science framework (Mullis et al., 2009). Several items and visualizations from the TIMSS 2011 study (see International Association for the Evaluation of Educational Achievement [IEA, 2013]) were also partly adapted to enhance the external validity of our results. Items confronted students with realistic situations, forcing them to

apply their basic science knowledge from biology ($n = 16$ items, e.g., food chains), physics ($n = 28$ items, e.g., gravity), and chemistry ($n = 4$ items, e.g., states of aggregation). All MC items comprised a short item stem with $M = 33.9$ words ($SD = 16.0$), a separate question, and four answer options (one correct; three distractors). The mean item word count was $M = 72.5$ ($SD = 25.1$). Each text-only item was experimentally manipulated by adding an RP to the item stem, visualizing the information from the text while never adding solution-relevant information that was unavailable in the text. RPs were realistic, schematic drawings in gray shades, displayed under the verbal item stem.

As reflected in the comparison of item examples in Figure 1, we constructed items in which the visualized item stem information was either *mandatory* ($n = 30$) or *optional* for answering the actual question ($n = 18$). This was done in order to systematically manipulate the RPs' task relevance, while nevertheless keeping the overall experimental framework comparable across all items. The manipulation was confirmed by three independent raters with a professional educational background. The raters' classification was unanimous for 44 items (92%); for the remaining four items, the majority opinion (i.e., the opinion shared by two raters) was used to classify the items. The whole test's EAP/PV reliability was estimated as .81 (based on a Rasch model), which can be considered very satisfactory.

General Cognitive Abilities

The non-verbal subtest N2 ("Figure Analogies"; adjusted according to students' grade; $\alpha = .91/.92$) of the Kognitiver Fähigkeitstest 4-12+ R (KFT; Heller & Perleth, 2000) was applied to measure spatial reasoning skills as an indicator of students' general cognitive abilities. The number of correct responses constituted the test score.

Reading Ability

We applied a German test from Retelsdorf, Becker, Köller, and Möller (2012) comparable to the German PISA decoding speed test that provides approximate information about basic reading skills (cf. Schneider, Schlagmüller, & Ennemoser, 2007). Within three minutes, students read a 740-word text that contained a total of 63 verbal numbers (e.g., thirty-one) that needed to be underlined while reading. Because it was not possible to finish the text within the allocated time, students were asked to highlight the last word they read. The number of words read (Read 1), the correctly underlined numbers (Read 2), and the missed numbers (Read 3) were applied as measures.

Test-Taking Pleasure

To measure students' affective state while working on the items, we had to apply an economical and easy-to-understand measure that did not seem strange to students when

https://econtent.hogrefe.com/doi/pdf/10.1027/1015-5759/a000351 - Thursday, April 25, 2024 7:08:23 AM - IP Address: 18.119.160.154

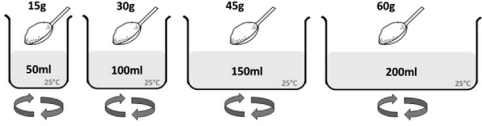
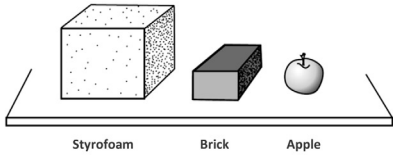
	Text-only item	Text-picture item
Mandatory-stem item	<p>Maria designed an experiment using salt and water. She experienced that 15g salt dissolved in 50ml water, 30g salt in 100ml water, 45g salt in 150ml water and 60g salt in 200ml water. The water's temperature was always 25°C. Maria stirred every mixture several times.</p> <p>What was Maria studying in her experiment?</p> <p>(A) How much salt will dissolve in different volumes of water. (B) How much salt will dissolve at different temperatures. (C) If stirring increases how fast salt will dissolve. (D) If stirring decreases how fast salt will dissolve.</p>	<p>Maria designed an experiment using salt and water. She experienced that 15g salt dissolved in 50ml water, 30g salt in 100ml water, 45g salt in 150ml water and 60g salt in 200ml water. The water's temperature was always 25°C. Maria stirred every mixture several times.</p>  <p>What was Maria studying in her experiment?</p> <p>(A) How much salt will dissolve in different volumes of water. (B) How much salt will dissolve at different temperatures. (C) If stirring increases how fast salt will dissolve. (D) If stirring decreases how fast salt will dissolve.</p>
Optional-stem item	<p>Jan's teacher places three objects on a table: A huge block of styrofoam, a normal brick and an apple. The styrofoam is bigger than the brick and the brick is bigger than the apple.</p> <p>Which statement is correct?</p> <p>(A) Objects with identical volume always weigh the same. (B) Heavy objects always have a greater volume than lighter objects. (C) Objects with a greater volume can be lighter than objects with a smaller volume. (D) Objects with a greater volume always weigh more than objects with a smaller volume.</p>	<p>Jan's teacher places three objects on a table: A huge block of styrofoam, a normal brick and an apple. The styrofoam is bigger than the brick and the brick is bigger than the apple.</p>  <p>Which statement is correct?</p> <p>(A) Objects with identical volume always weigh the same. (B) Heavy objects always have a greater volume than lighter objects. (C) Objects with a greater volume can be lighter than objects with a smaller volume. (D) Objects with a greater volume always weigh more than objects with a smaller volume.</p>

Figure 1. Examples of experimentally manipulated text-only and text-picture science items (material adapted from IEA, 2013) with a mandatory or an optional item stem.

assessed for every single item. Thus, we came up with a one-item measure, asking students whether they had fun solving the current item (“Working on this item was fun for me.”). Because the rating of the item was of primary interest here, the focus was explicitly placed on the item and not on the student working on the item (e.g., asking for motivation or engagement). Furthermore, students rated their general attitude towards the pictures (“liking” them). Again, all measures were assessed on a 4-point Likert scale.

Test-Effort

Students indicated, in a measure from PISA (*Effort Thermometer*; Kunter et al., 2002), how much effort they planned to invest in the (following) science test compared to a maximum effort in a situation of high personal importance.

Procedure

Test administrators conducted the study at schools during lessons. After an introduction, students answered the background questionnaire, took the reading test, and answered the Effort Thermometer. Students were instructed in written and oral form on how to answer the science items,

which included always providing an answer and weighted guessing in cases of doubt. For each item, students rated their enjoyment (“fun”) when solving it. The science test was administered without time restrictions. After the science test, students evaluated the pictures and worked on the KFT, after receiving standardized advice.

Analyses

Following the multimatrix design, we applied IRT models that can handle the large amount of missing values by design when analyzing performance data. As a rough indicator for how appropriate the constructed items were for the sample and the following IRT analyses, we evaluated Rasch-estimated item difficulties and corresponding weighted mean square (WMNSQ) indices.

We followed an exploratory and theory-driven approach, considering different IRT models to conceptualize the generalizability and effect size of the MME(s) across test items and students, and to understand how much the relevance of the information visualized by RPs (*mandatory* vs. *optional*) impacted the MME. We used Linear Logistic Test Models (LLTM; Fischer, 1973) and Latent Difference Score

(LDS) models (e.g., Steyer, Eid, & Schwenkmezger, 1997) to specify eight competing models that comprised a latent “science” factor and different MME conceptualizations (see Figure 2).

A first LLTM (M1) assumed identical difficulties for text-picture and text-only items, which equal the complete absence of an MME (i.e., null hypothesis). Unrestricted difficulty estimation for each text-only and text-picture item in a one-dimensional Rasch model (M2) implied that item-specific MMEs exist. Another LLTM (M3) assumed one overall MME across all items. A third LLTM (M4) considered two separate MMEs for item stems with mandatory and optional information. Furthermore, another LLTM (M5) defined an MME for mandatory-stem items but no MME for optional-stem items (identical difficulties for text-picture and text-only items). In a two-dimensional LDS model (M6), one overall MME was specified for all items as a latent difference factor that can vary across students. A three-dimensional LDS model (M7) included two separate MMEs as latent difference factors for optional- and mandatory-stem items that can both vary across students. Finally, a last two-dimensional LDS model (M8) was specified with an MME only for mandatory-stem items (varying across students) and not for optional-stem items. The fit of all the models was compared using the Bayesian information criterion (BIC). Building on the best model candidate, we used an explanatory model (M9) to realize a regression of individual MME(s) on students’ characteristics. To prevent highly correlated predictors from depressing each other in this approach, we first identified four groups of fairly lowly correlated predictors by running a factor analysis. Accordingly, we specified four separate explanatory LDS models (M9 A-D) to estimate the predictive power of students’ characteristics for individual MME(s).

To analyze students’ ratings of item-solving pleasure, we used a Rating Scale Model (Andrich, 1978) to account for the missing values by design and a paired *t*-test to inferentially compare the resulting pleasure parameters for text-only and text-picture items.

We used IBM® SPSS 19 for the statistical analyses and *Mplus* Version 7.3 (Muthén & Muthén, 1998–2015) for the psychometrical analyses.

Results

Thirty-six to 40 students ($M = 38$; $SD = 1.55$) worked on each of the eight booklets, distributed equally across school tracks, with 68–76% ($M = 73\%$; $SD = 2\%$) from academic and 24–33% ($M = 27\%$; $SD = 2\%$) from nonacademic track

schools per booklet. Each item was solved by at least $n = 99$ students ($M = 102.5$; $SD = 2.56$). Item solution frequency ranged from .15 to .97 ($M = .65$; $SD = .18$). Apart from 62% missings by design, no missing values appeared in the science test.

Item Parameters

The difficulty parameters obtained from a Rasch model (M2) reflected that the items were rather easy while still being appropriate for the sample ($M = -0.794$; $SD = 1.09$; range = -3.78 – 2.11). Most test items showed a good fit according to their WMNSQ indices ($M_{WMNSQ} = 0.99$; $SD_{WMNSQ} = 0.12$); only 5.2% of the items did not fit the Rasch model (cf. ESM 2).

We compared the model fit indices of the eight competing IRT models (see Figure 1 and Table 1) to investigate the existence of the expected decrease in item-difficulty caused by RPs (MME), and its generalizability across items (specifically across mandatory-stem and optional-stem items) and across students. The model with the overall best fit (lowest BIC) was model M5, with very strong evidence towards this model in comparison to the other models, shown in the BIC differences (cf. rule of thumb¹; Raftery, 1999) which were 10 points or more. Only one smaller BIC difference (5.7 points) was observed – in comparison to model M4, which nevertheless reflected clear positive evidence in favor of model M5.

The MME (i.e., medium item-difficulty decrease) caused by RPs for mandatory-stem items was estimated as $M = -0.417$ ($SE = 0.062$) in model M5 with a medium effect size of $d = .53$. The MME for items with an optional stem was fixed to zero, indicating that an overall effect was probably nonexistent or extremely small. Model M5 also suggested that the effect for mandatory-stem items was a constant, as the MME did not vary across students. To visualize changes in item difficulties caused by RPs, Figure 3 shows the item-specific MME for all 48 parallel items, as well as the mean MME estimation (for mandatory-stem items) following the preferred model (M5).

Because optional-stem items had much fewer words than mandatory-stem items ($M_{MAN} = 77.5$, $SD_{MAN} = 4.5$; $M_{OPT} = 64.1$, $SD_{OPT} = 5.6$), we checked whether the number of words interacted with the MME, as such an interaction could be an alternative explanation for the difference in MME estimates for mandatory- and optional-stem items. Therefore, we computed the correlation between the number of words per item with the item-specific difference between text-only and text-picture difficulty estimates (drawn from the Rasch model: M2).

¹ Raftery (1999) suggested interpreting BIC differences between 0 to 2 as *weak*, 2 to 6 as *positive*, 6 to 10 as *strong*, and greater than 10 as *very strong* evidence towards a model.

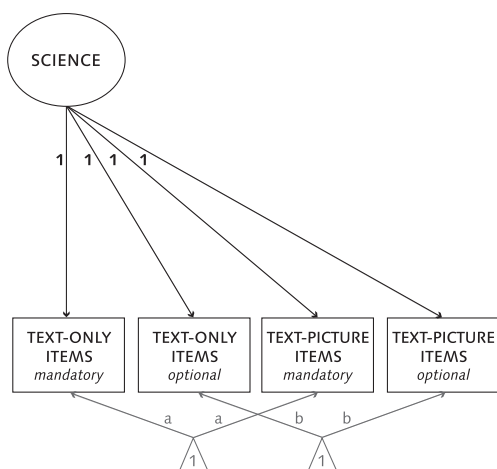
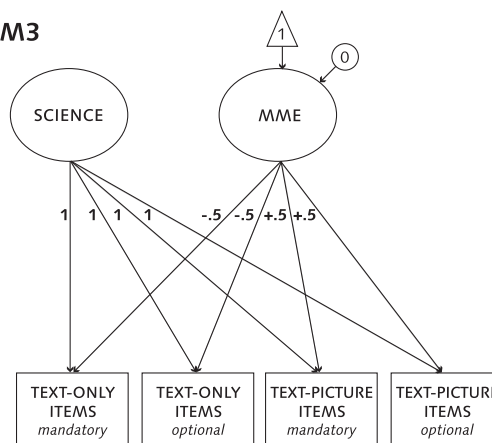
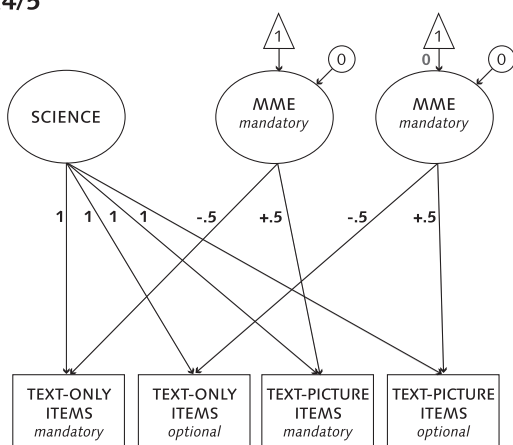
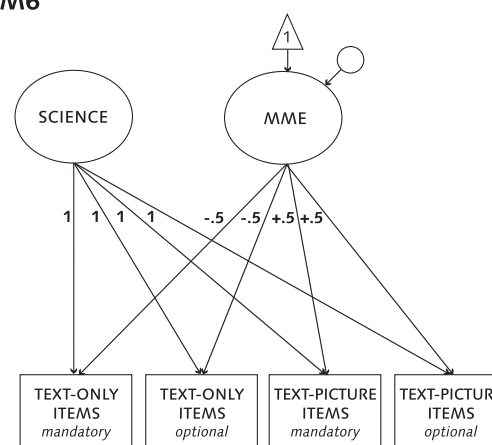
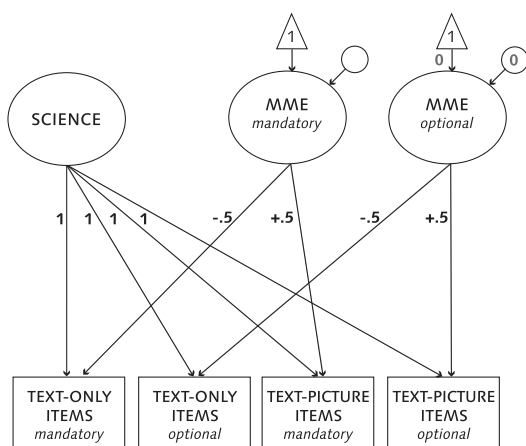
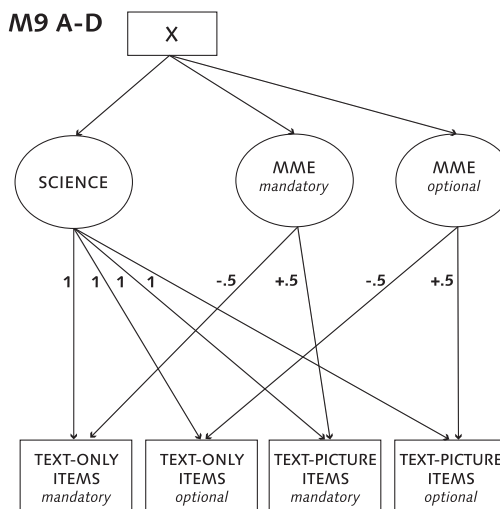
M1/2**M3****M4/5****M6****M7/8****M9 A-D**

Figure 2. Models applied to specify different MME conceptualizations, descriptions are provided in the text. Gray parts indicate minor changes between two related models that are displayed in one schema.

Table 1. Results of the eight model comparisons

Model	# Parameters	N	Deviance	AIC	BIC	BIC _{Adj}
M1	49	305	10,680.48	10,778.47	10,960.77	10,805.37
M2	97	305	10,547.88	10,741.88	11,102.75	10,795.11
M3	50	305	10,650.92	10,750.91	10,936.93	10,778.35
M4	51	305	10,635.18	10,737.17	10,926.91	10,765.16
M5	50	305	10,635.18	10,735.18	10,921.20	10,762.62
M6	52	305	10,649.10	10,753.10	10,946.55	10,781.63
M7	56	305	10,630.96	10,742.96	10,951.30	10,773.70
M8	52	305	10,634.32	10,738.32	10,931.78	10,766.86

Notes. Model descriptions (M1–M8) are provided in the text; AIC = Akaike information criterion; BIC = Bayesian information criterion; BIC_{Adj} = Sample size adjusted BIC.

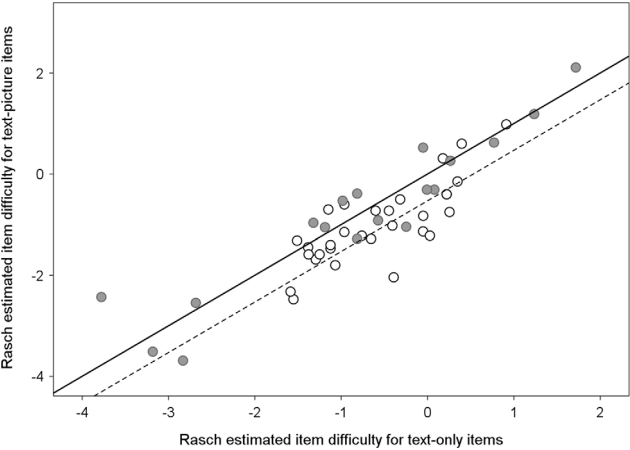


Figure 3. Item difficulty shift for parallel text-only and text-picture items (cf. Rasch model M2). Dots below the (black) origin line represent a decrease in difficulty. Based on the best model (M5), the origin line also represents the overall zero MME for optional-stem items (gray circles), while the dashed line represents the MME effect size for mandatory-stem items (white circles).

This correlation turned out to be almost zero ($r = -.08$; $p = .59$), suggesting that the lower word count was probably not a reason for the missing MME in optional-stem items.

Students' Characteristics

Although the residual variance of the MME was fixed to zero in the best fitting model (M5), we applied the model for the planned explanatory regression approach in order to explore whether interactions stood out between students' characteristics and the MME in mandatory-stem items. In this model (M9), variance was assigned to the MME score by the explanative power of the predictors, while the residual variance was still fixed to zero. The standardized regression coefficients of the explanatory LDS models (M9, A–D) are displayed in Table 2. It is

noteworthy that the standardized variance component predicted by students' characteristics always equaled one, which influenced the values of the regression coefficients. As was to be expected from the model fit improvements whenever the variance of the MME (overall and for mandatory-stem items) was fixed to zero, there was not much between-student variation to explain, while the existing MME variation could hardly be predicted by students' characteristics. Significant regression weights ($p < .001$) were only found for students' *age* and their *overall test-taking pleasure*. All other predictors were nonsignificant (cf. Table 2).

Test-Taking Pleasure

The *t*-test revealed that students found solving text-picture items to be significantly more fun than solving parallel text-only items ($M_{\text{text-only}} = -0.03$, $SD = 0.32$; $M_{\text{text-picture}} = -0.24$, $SD = 0.32$; $t(47) = 5.71$, $p < .001$), while effect sizes for mandatory-stem items were higher ($d = 1.00$) than for optional-stem items ($d = 0.71$). Furthermore, students' rating of liking the pictures ($M = 3.12$; $SD = 0.95$) was significantly above the scale average ($p < .001$) and above category '3' (= *agree*; $p = .02$) on the 4-point Likert scale.

Discussion

Given the frequent use of pictures in standardized assessment, it is important to understand how item characteristics and students' characteristics affect the impact of pictorial representations on test outcomes and on students' test-taking experience. Following an experimental within-subject IRT approach, the present study aimed to provide new insights into this issue for the case of RPs in science assessment, thereby making a contribution towards the

Table 2. Means (*M*) or frequencies, standard deviations (*SD*), and standardized regression coefficients β from the explanatory Latent Difference Score Models A–D (M9)

	<i>M</i>	<i>SD</i>	Model A	Model B	Model C	Model D
Age	11.55	0.77				–.93***
Gender ^b	47.20	–	.26			
School track ^b	72.50	–	.55			
KFT _(max. 25)	18.05	6.24		–.67		
Read 1(# words read, max. 740)	350.9	96.03	–.20			
Read 2(# underlined numerals, max. 63)	27.56	8.22		.35		
Read 3(# missed numerals, max. 63)	1.67	2.84			.62	
Reading self-concept _(max. 16)	12.21	2.45				.36
German grade _(max. 6)	2.56	0.79			.58	
Science grade _(max. 6)	2.38	0.76				.40
Science interest _(max. 20)	12.04	3.66			.53	
Test-taking pleasure	0 ^a	2.11	.82***			
Planned test-effort _(max. 10)	8.08	1.91		.79		

Notes. *** $p < .001$; ^aEstimated mean was restricted to zero in the IRT estimation procedure; ^bFor dummy coded variables, instead of the mean, the frequency of category '1' is reported in percent: *Gender* (0 = male; 1 = female); *School track* (0 = regional school; 1 = Gymnasium).

recently developed body of research in this field (e.g., Hartmann, 2012; Prenzel et al., 2002; Saß et al., 2012; Wu, Kuo, Jen, & Hsu, 2015).

First, we were able to replicate an item difficulty decrease caused by adding RPs (e.g., Saß et al., 2012), while the comparison of several competing IRT model conceptualizations indicated that this MME in testing depends on the relevance of item-stem information and the relevance of the RP for answering items correctly. Accordingly, *mandatory-stem items* showed a medium ($d = .52$) decrease in item difficulty caused by the integration of an RP, while our findings suggest that this decrease can be considered to be constant across students. At the same time, the best model conceptualization implied that there was no MME for *optional-stem items*, indicating that RPs that provided context information that was not essential for solving the actual question neither helped nor harmed students' performance. This is perhaps comparable to findings from Wu et al. (2015), which showed that static compared to dynamic graphics in test items affected students' performance differently, but only when pictures displayed mandatory compared to optional information. In the present study, an alternative explanation for the absent MME in optional-stem items based on a lower word count turned out to be implausible. Still, the impact of text length needs to be investigated further because our items had short texts with little variance. This could have caused an underestimation of the influence of the text length on the MME. However, data from mandatory-stem items show that the MME does not necessarily depend on replacing a longer text with a picture (cf. Hartmann, 2012), but rather on forming a better understanding of the problem by simultaneously using related verbal and pictorial information to build a coherent mental model

(Schnotz & Bannert, 2003), even in short items. Overall, these results indicate that cognitive facilitations caused by RPs are closely related to students building a situational mental model relevant to the requested task.

Second, IRT model comparisons indicated that the MME in mandatory-stem items can be considered a stabile phenomenon that affects all students' performance in a comparable manner. This implies that effects on item difficulty caused by adding RPs are relatively predictable, while RPs seemingly do not change the measurement of the construct, which would be preferable in terms of construct validity. This also indicates that not much interaction can be expected at an individual student level: We found no evidence that students' *gender*, *school track*, *general cognitive abilities*, *reading abilities* (Read 1–3), *reading self-concept*, *grades*, *planned test-effort*, or *science interest* were predictive of the impact that RPs had on their test performance (i.e., the MME in testing). Thus, in contrast to findings by Hartmann (2012) for items with longer texts, we were not able to replicate a differential effect of RPs for poor readers. This could mean that RPs do not compensate for poor reading ability in items with a short text that has been complemented and not replaced by the RP. But it could also be connected to the accuracy of the applied reading test. Given the potential relevance for item construction, future research should reconsider reading effects in items with different text length, using more sophisticated (e.g., computerized) reading measures.

However, students' *age* and *overall test-taking pleasure* were identified as significant predictors of the MME in mandatory-stem items. Considering the predictors' algebraic signs, younger students seemingly benefited more from RPs, while students with higher test-taking pleasure also showed a higher MME. This might indicate that RPs'

support for students in building a mental model was even more important for younger students. The association with the overall test-taking pleasure might relate to a cognitive-affective mediation effect (e.g., Moreno & Mayer, 2007). Accordingly, students with higher motivational levels might have put more effort into integrating the text and picture information into a coherent situational mental model (cf. Mayer, 2005; Schnotz & Bannert, 2003), resulting in a higher MME. However, these effects need to be interpreted with great care, because this analysis might not be completely reliable as it was based on a model with fixed residual variance. Altogether, our data provide clear evidence that the MME is not a selective phenomenon, affecting students differently according to their individual characteristics, but rather a calculable constant.

Third, this study provides empirical evidence for the assumption that pictures enhance students' test-taking pleasure when solving text-picture compared to text-only items. Students also clearly indicated that they liked the pictures in the test. However, the task relevance of RPs moderated the effect, reflected in a strong effect size ($d = 1.00$) for mandatory-stem items and a considerable but lower effect size for optional-stem items ($d = 0.71$). Thus, students enjoyed solving items even more when the RP displayed information that was essential for solving the task, while optional RPs still substantially raised students' test-taking pleasure. These findings are highly relevant for test construction in the context of low-stakes assessment (cf. Wise & DeMars, 2005). Our results are also in line with research from multimedia learning, which showed that decorative pictures (similar to optional RPs) had a positive effect on students' learning enjoyment even though they did not affect performance (Lenzner et al., 2013). Furthermore, according to studies on test-taking motivation, students are generally more likely to show solution behavior when item formats are perceived as being less complex (e.g., Asseburg & Frey, 2013; Wise & DeMars, 2005). This might explain students' positive attitudes towards the pictures in general and their implicit preference for RPs that provide mandatory rather than optional information. This relates to the abovementioned fact that essential RPs help students to build an adequate problem-related mental model with less effort in less time (Eitel et al., 2013; Mayer, 2005; Schnotz & Bannert, 2003), thereby reducing the mental complexity while working on the presented problem. Overall, increasing students' test-taking pleasure might also mean increasing their task engagement; this could be reflected in reduced frequencies of rapid-guessing behavior, for example, as the observation from Wise et al. (2009) suggests. However, research that uses more direct parameters, such as reaction-time or eye-movement measures while students solve test items

(cf. e.g., Lindner et al., 2014), is needed to support this claim empirically.

Taken together, the results of the present study show that the task relevance of RPs not only moderates positive effects on students' performance, but also influences the extent to which RPs increase students' test-taking pleasure. Nevertheless, as RPs (whether task-relevant or not) enhanced students' pleasure to a substantial degree, the present findings might encourage the integration of task-relevant RPs or even decorative pictures into low-stakes assessment in order to improve students' affective state when solving items. This suggestion is further supported by our findings that the MME was a predictable effect that either influenced item difficulty (when RPs displayed task-relevant information) or did not (when RPs displayed optional context information). Although more research is clearly still necessary, test constructors should seriously consider the choice of pictures in item writing.

Acknowledgments

We thank Julia Barenthien and Benjamin Strobel for helping with data collection.

Electronic Supplementary Material (ESM)

The electronic supplementary material is available with the online version of the article at <http://dx.doi.org/10.1027/1015-5759/a000351>

ESM 1. Text and Table (PDF).

Comments on the experimental multimatrix design (ESM1_Multimatrix_design.pdf).

ESM 2. Text and Table (PDF).

Supplementary material regarding the measures in the study (ESM2_Measures.pdf).

References

- Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education*, 33, 131–152. doi: 10.1016/S0360-1315(99)00029-9
- Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 2, 581–594. doi: 10.1177/014662167800200413
- Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, 55, 92–104.
- Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Clarendon Press. doi: 10.1002/acp.2350020209
- 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. doi: 10.1016/j.learninstruc.2013.05.002

- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359–374. doi: 10.1016/0001-6918(73)90003-6
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15, 309–344. doi: 10.1207/S15324818AME1503_5
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.
- 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/DocumentServlet/33260/hartmann_diss.pdf
- Heller, K. A., & Perleth, C. (2000). KFT 4–12+ R: *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision* [Cognitive Abilities Test for students from grade 4 to 12+ (CogAT; Thorndike, L. & Hagen, E., 1954–1986) German adapted version/author]. Göttingen, Germany: Beltz.
- IEA [International Association for the Evaluation of Educational Achievement]. (2013). *TIMSS 2011 assessment released science items*. Retrieved from http://nces.ed.gov/timss/pdf/TIMSS2011_G4_Science.pdf. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College
- Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., ... Weib, M. (2002). *German scale handbook for PISA 2000*. Berlin, Germany: Max-Planck-Institut für Bildungsforschung.
- Lenzner, A., Schnotz, W., & Müller, A. (2013). The role of decorative pictures in learning. *Instructional Science*, 41, 811–831. doi: 10.1007/s11251-012-9256-z
- Levie, H. W., & Lentz, R. (1982). Effects of text illustration: A review of research. *Educational Communication & Technology Journal*, 30, 195–232. doi: 10.1007/BF02765184
- Lindner, M. A., Eitel, A., Thoma, G.-B., Dalehefte, I. M., Ihme, J. M., & Köller, O. (2014). Tracking the decision making process in multiple-choice assessment: Evidence from eye movements. *Applied Cognitive Psychology*, 28, 738–752. doi: 10.1002/acp.3060
- Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mayer, R. E. (Ed.). (2005). *The Cambridge handbook of multimedia learning*. Cambridge, UK: University Press. doi: 10.1017/CBO9780511816819.005
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
- Möller, J., & Bonerat, E. M. (2007). Fragebogen zur habituellen Lesemotivation. [Habitual reading motivation questionnaire]. *Psychologie in Erziehung und Unterricht*, 54, 259–267.
- Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, 19, 309–326. doi: 10.1007/s10648-007-9047-2
- Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).
- Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- OECD. (2009). *PISA 2006 Technical Report*. Paris, France: OECD Publishing. doi: 10.1787/9789264167872-en
- OECD. (2013). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris France: OECD Publishing. doi: 10.1787/9789264190511-en
- Paivio, A. (1986). *Mental representations: A dual coding approach*. New York, NY: Oxford University Press.
- Prenzel, M., Häußler, P., Rost, J., & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? [The PISA science literacy test: Are the item difficulties predictable?]. *Unterrichtswissenschaft*, 30(2), 120–135. doi: nbn:de:0111-opus-76826
- Raftery, A. E. (1999). Bayes factors and BIC: Comment on “A critique of the Bayesian information criterion for model selection”. *Sociological Methods & Research*, 27, 411–417. doi: 10.1177/0049124199027003005
- Retelsdorf, J., Becker, M., Köller, O., & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *British Journal of Educational Psychology*, 82, 647–671. doi: 10.1111/j.2044-8279.2011.02051.x
- 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. doi: 10.1002/acp.1798
- Schneider, W., Schlagmüller, M., & Ennemoser, M. (2007). LGVT 6–12: *Lesegeschwindigkeits- und -verständnistest für die Klassen 6–12* [LGVT 6–12: A reading comprehension test for students from grade 6 to 12]. Göttingen, Germany: Hogrefe.
- Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, 13, 141–156. doi: 10.1016/S0959-4752(02)00017-8
- Schnotz, W., Ludewig, U., Ullrich, M., Horz, H., McElvany, N., & Baumert, J. (2014). Strategy shifts during learning from texts and pictures. *Journal of Educational Psychology*, 106, 974–989. doi: 10.1037/a0037054
- Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online*, 2, 21–33.
- Vekiri, I. (2002). What is the value of graphical displays in learning? *Educational Psychology Review*, 14, 261–312. doi: 10.1023/A:1016064429161
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10, 1–17. doi: 10.1207/s15326977ea1001_1
- 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. doi: 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. doi: 10.1016/j.compedu.2015.01.007

Received July 11, 2015

Revision received February 2, 2016

Accepted February 9, 2016

Published online October 7, 2016

Marlit Annalena Lindner

Leibniz Institute for Science and Mathematics Education
Olshausenstraße 62
24118 Kiel
Germany
Tel. +49 431 880 4410
Fax +49 431 880 2629
E-mail mlindner@ipn.uni-kiel.de