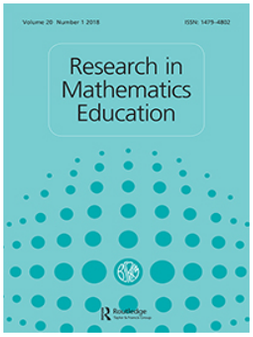

Research in Mathematics Education

ISSN: 1479-4802 (Print) 1754-0178 (Online) Journal homepage: https://www.tandfonline.com/loi/rrme20

Word problems versus image-rich problems: an analysis of effects of task characteristics on students' performance on contextual mathematics problems

Kees Hoogland, Birgit Pepin, Jaap de Koning, Arthur Bakker & Koeno Gravemeijer

To cite this article: Kees Hoogland, Birgit Pepin, Jaap de Koning, Arthur Bakker & Koeno Gravemeijer (2018) Word problems versus image-rich problems: an analysis of effects of task characteristics on students' performance on contextual mathematics problems, Research in Mathematics Education, 20:1, 37-52, DOI: 10.1080/14794802.2017.1413414

To link to this article: https://doi.org/10.1080/14794802.2017.1413414

RESEARCH IN MATHEMATICS EDUCATION, 2018
VOL. 20, NO. 1, 37–52
https://doi.org/10.1080/14794802.2017.1413414

Word problems versus image-rich problems: an analysis of effects of task characteristics on students' performance on contextual mathematics problems

Kees Hoogland [a], Birgit Pepin[b], Jaap de Koning[c], Arthur Bakker[d] and Koeno Gravemeijer[b]

[a]Research Centre for Learning and Innovation, Utrecht University of Applied Sciences, Utrecht, The Netherlands; [b]Eindhoven School of Education, Eindhoven University, Eindhoven, The Netherlands; [c]SEOR, Erasmus University Rotterdam, Rotterdam, The Netherlands; [d]Freudenthal Institute, Utrecht University, Utrecht, The Netherlands

ABSTRACT
This article reports on a *post hoc* study using a randomised controlled trial with 31,842 students in the Netherlands and an instrument consisting of 21 paired problems. The trial showed a variability in the differences of students' results in solving contextual mathematical problems with either a descriptive or a depictive representation of the problem situation. In this study the relation between this variability and two task characteristics is investigated: (1) complexity of the task representation; and (2) the content domain of the task. We found indications that differences in performance on descriptive and depictive representations of the problem situation are related to the content domain of the problems. One of the tentative conclusions is that for depicted problems in the domain of measurement and geometry the inferential step from representation of the problem situation to the mathematical problem to be solved is smaller than for word problems.

ARTICLE HISTORY
Received 12 February 2016
Accepted 25 November 2017

KEYWORDS
contextual mathematics problem; task characteristics; word problem

Introduction

This study is part of a larger research project to investigate alternatives for the persistent and problematic use of word problems to teach and assess students' ability to deal with numerical problems based on everyday life situations. In current classroom practice word problems are predominantly used to teach and assess these abilities, although many research findings over the past 20 years have reported serious difficulties in using word problems for this purpose (Verschaffel, Greer, & De Corte, 2000; Verschaffel, Greer, Van Dooren, & Mukhopadhyay, 2009). We expected that the use of images from real life would counteract these difficulties. Therefore, we designed tasks that were more authentic by changing the representation of the problem situation from descriptive to more depictive.

CONTACT Birgit Pepin b.e.u.pepin@tue.nl

In an earlier study (Hoogland, Pepin, Bakker, de Koning, & Gravemeijer, 2016) we showed in detail the design and validation of the instrument with which we measured the effect of changing the representation of the problem situation on students' performance in a randomised controlled way. This instrument is explained in the method section. Using this instrument, we found in the overall test result (Hoogland, 2016) that the students performed significantly better on image-rich problems than on otherwise equivalent word problems, with an increase of correct responses of about two percentage points (Hoogland, 2016). This result was a general overall result, in the sense that aggregated students' results were analysed. In the aggregated result the measured effect was independent of background variables such as age of students, ability of students, ethnicity and (type of) school and a small interdependency with gender was measured.

In the original study we observed that the effect of changing the representation varied considerably between the pairs of problems. To get a better understanding of the underlying patterns, in the current study we focus on this variability. We analyse *post hoc* the interdependence of this variability with two task characteristics: (1) complexity of the task representation, measured in number of words or number of pictures; and (2) the content domains of the tasks.

Theoretical and empirical background

Over the past decades problem-solving has gained importance as one of the main goals of mathematics education (Schoenfeld, 2014). For both teaching and assessment purposes, problem-solving tasks have been designed by mathematics educators worldwide, especially in the field of numeracy and mathematical literacy (Geiger, Goos, & Forgasz, 2015). The aim of these problem-solving tasks is that students solve the posed problems using their mathematics knowledge and skills.

Until recently, the predominant representation of a problem situation has been verbal (Verschaffel et al., 2009), the typical consequence of using pen and paper, a typewriter or a simple word-processor to design tasks. This has led to the genre of "word problems". According to Verschaffel, Depaepe, and Van Dooren (2014) word problems can be defined as "verbal descriptions of problem situations wherein one or more questions are raised the answer to which can be obtained by the application of mathematical operations to numerical data available in the problem statement" (p. 641). In word problems both the description of the problem situation as well as the actual problem statement are presented in words. The predominance of the verbal description of problem situations can most clearly be demonstrated by the intriguing fact that the concept word problem in education almost exclusively means a contextual mathematical problem.

There are many studies that report serious difficulties with the use of word problems in the mathematics classroom. The most reported difficulty is that students do not take into account common sense considerations about the problem (Greer, 1997; Verschaffel, Corte, & Lasure, 1994), which affects the processes of formulating the mathematical problem, and interpreting the mathematical results. As a consequence, many students fail to solve the posed problems correctly. Common in many analyses of the difficulties encountered is that students look at these problems with a strong "answer-getting mind-set" (Daro, 2013) and that they have a calculational approach to mathematics, often related to the calculational approach used by their teachers (Thompson, Philipp, Thompson, & Boyd, 1994). This "answer-getting mind-

set" is arguably a result of the mind-sets of both students and teachers (Depaepe, De Corte, & Verschaffel, 2010). There are persistent socio-mathematical norms (Gravemeijer, 1997; Yackel & Cobb, 1995) in many mathematics classrooms implying that solving problems of any kind means "getting the right answer" by conducting a series of operations on the numbers in the problem without making sense of the situation. This so-called "suspension of sense-making" (Schoenfeld, 1992) leads to mistakes by students and as a consequence to underachievement of students. In our research we searched for alternatives to word problems which were expected to strengthen the association with real world situations (Palm, 2009) and therefore decrease the suspension of sense-making and the strong calculational focus (Thompson et al., 1994). As an indicator for such reduction we investigated the change in students' results. We adopted the approach of systematically replacing verbal representations of the problem situations by depictive ones, with the idea that depictive representations of the problem situations stay closer to the real problems that are represented, and that students are more likely to make sense of a pictorial situation. Subsequently, they may more likely adopt a problem-solving attitude, with more chances of solving the posed problem correctly.

A study by Dewolf et al. was of special relevance to our design (Dewolf, Van Dooren, Ev Cimen, & Verschaffel, 2014; Dewolf, Van Dooren, Hermens, & Verschaffel, 2015). They have reported on the effect of adding decorative, representational, and informational pictures in contextual mathematical problems: they found hardly any effects on student behaviour in solving the problem, nor in reflecting on the correctness of their solution. However, they presented the illustrations next to the texts of the word problems, rather than trying to integrate text and picture, stipulating this could explain the absence of effect on student behaviour. This study encouraged us to design problems, where the authentic images are integrated in the problem representation, to see if this could generate more effect on student behaviour.

Although the prediction of higher scores when a depictive representation of the problem situation was used, was confirmed in an earlier study (Hoogland, 2016), the results on the task level were not straightforward. In the current study we pursued a *post hoc* analysis of these findings on task level to shed more light on the intricate relation between task characteristics and student performance. We emphasise that these characteristics were not manipulated in the aforementioned experiment but studied now in retrospect as potential explanations of differences found in performance.

The complexity of the tasks

Several different perspectives and frameworks can be used to define task complexity in mathematics education (Watson & Ohtani, 2015; Williams & Clarke, 1997) and in general educational settings (Gill & Hicks, 2006; Robinson, 2001).

In the case of contextual problems, "wordiness" is often used in research on task design, for example, by Piel and Schuchart (2014), albeit anchored in general research on the effect of text comprehension on students' performance. This research claims that wordiness often has a negative effect on students' performance (Boonen, van Wesel, Jolles, & van der Schoot, 2014; Fuchs, Fuchs, Compton, Hamlett, & Wang, 2015). We argued, likewise, that high wordiness would predict a negative effect on students' performance, and therefore would strengthen the positive effect on students' results of changing the representation of the problem situation from descriptive to more depictive.

In line with the definition of wordiness, we defined complexity of the image-rich problems as the number of visual elements used. This approach is commonly used in research on cognitive load theory on element interactivity (Leppink, Paas, van Gog, van der Vleuten, & van Merriënboer, 2014). In this research a high number of visual elements is reported to have a negative impact on students' performance. We argued likewise, and predicted that a high number of visual elements would limit the effect of changing the representation of the problem situation from descriptive to more depictive.

Some depictive representations of the problem situation also contain some words due to the nature of the context, for instance a recipe. We argue that words to describe a situation are another category and have another function than words that are inherent to the context of the problem. For the current analysis we disregarded this kind of word use. So, in our statistical model we took wordiness for the descriptive version tasks, and the number of visual elements for the depictive version of the tasks as additional independent variables.

The content domains of the tasks

We further assumed that the difference in students' performance on the two versions of the paired problems may work out differently on tasks in different content domains. In the Netherlands in 2010 a new Numeracy Framework (Hoogland & Stelwagen, 2011) was introduced as part of the "Referentiekader Taal en Rekenen" (Literacy and Numeracy Framework [LaNF]). The content domains in this framework are numbers, proportions, measurement & geometry, and relations. The Dutch framework was inspired by international frameworks of numeracy and mathematical literacy, such as TIMSS, PISA and PIAAC (Mullis & Martin, 2013; OECD, 2013; PIAAC Numeracy Expert Group, 2009), albeit with a stronger focus on numbers and proportions rather than probability and statistics.

We argue that the mental activity needed for the necessary steps in the problem-solving process is dependent on the mathematics domain of the task (Schnotz, Baadte, Müller, & Rasch, 2010). In the domain of numbers, the mathematical model is primarily computational and thus one-dimensional. In that case a more depictive representation was presumed not to contribute considerably to the ease with which problem-solvers make the situational or mathematical model. However, in the domain of proportions the mathematical model is in general more complex than in the domain of numbers because there is always some activity of relatively comparing quantities or comparing a quantity to a whole and relatively comparing is considered to be more abstract than straightforward operations. A more depictive representation was assumed to be beneficial here to support the relative thinking in proportions. At the same time a counter-effect seems possible if the mental model and the depictive representations are not mutually beneficial. In that case, an addition of complexities is likely to be experienced by the students. Hence, for tasks from the domain of proportions one could not make a plausible straightforward prediction, whether a more depictive representation could help the solvers to construct the appropriate mental model and hence help them in solving the problem in a successful way.

In the domain of measurement & geometry the underlying problem situation is often in itself two- or three-dimensional. Hence, a more depictive representation of the problem was assumed to most likely support the problem-solver to create the appropriate mental and mathematical model.

Research method

Description of the used instrument

The instrument

The instrument used to collect the data was a web-based numeracy test of 21 tasks (Hoogland et al., 2016). These were evenly spread over the content domains of numbers, proportions, and measurement & geometry and very similar in content and layout to the nationwide examination (Cito instituut voor toetsontwikkeling, 2015).

For each individual run half of the 21 tasks (10 or 11) were randomly chosen to be presented as word problems (A-version), and the others as image-rich numeracy problems (B-version). Furthermore, for each individual run the tasks were presented in random order. Table 1 provides an overview of the tasks used in the trial, supplemented with the quantification of the characteristics which we added as variables for this study. Figure 1 shows two examples of paired problems used in the instrument, translated in English for readability. The data in Table 1 were derived from the original Dutch versions of the tasks. The coding was done independently by researchers, and agreement on codes was established after discussion. For instance, we counted numbers as words, and we counted yellow Post-its™ as 1 pictorial element. The Dutch version and the English translation can be found under open access (Hoogland & De Koning, 2013). In our statistical model we took data on the background of the participants into account, such as gender, ethnicity, school level, and mathematics level. By adding other characteristics of the used problems, such as complexity and cognitive/content domains, we were able to shed light on the interdependences of, for instance, problem complexity and effect of changing the representation of the problem situation.

Table 1. Task characteristics of 21 paired problems.

Domain	Task	Item	Wordiness	Pictorial Complexity
Numbers	TV + dvd	4	26	2
	Change	5	19	2
	Money pile	9	36	2
	Kitchen tiles	12	35	2
	Hamburgers	16	21	2
	Coughing syrup	17	36	3
	Public debt	18	18	2
Proportions	Travel time	3	29	2
	Recipe	6	44	1
	Price magazine	7	36	2
	AEX index	8	30	3
	Scale model	10	23	2
	Endive	15	9	1
	Winter tires	20	40	3
Measurement & Geometry	Apples in bag	1	23	2
	Gas usage	2	30	2
	Double glazing	11	27	2
	Water bottles	13	19	2
	Bedroom tiles	14	38	3
	Cake tin	19	17	1
	Chocolate boxes	21	51	4

Note: Wordiness is number of words used in the description of the problem situation (A-version, Dutch version). Pictorial Complexity is the number of visual elements in the depictive representation (B-version, Dutch version).

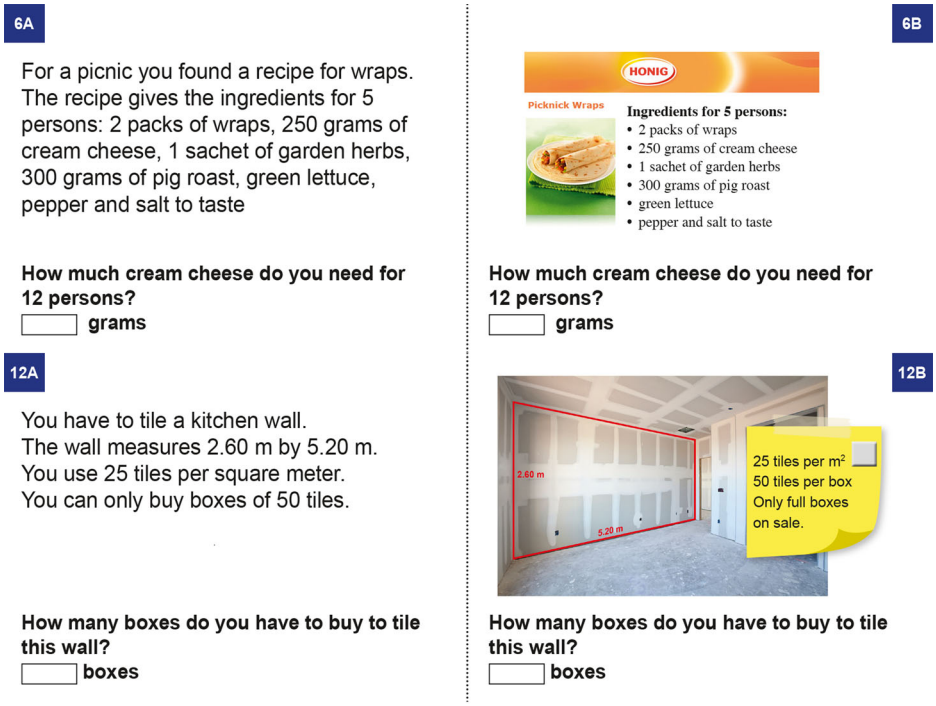

Figure 1. Two examples of paired problems: item 6 and item 12.

The participants

The test was made available through internet for all schools nationwide to use as a diagnostic test. In total, 31,842 students from 179 schools geographically spread across the Netherlands, participated in the test (See Table 2 for school level and age group). In terms of the total student population in the Netherlands (CBS, 2012) around 2% participated.

The distribution of participants over gender and ethnicity were close to the national percentages (CBS, 2012). In the distribution over the various school levels secondary students were over-represented but in all school levels the patterns in the results were the same (Hoogland, 2016).

Statistical analysis

For this study a statistical analysis was carried out on the separate couples of paired problems. The analysis focused on the differences in scores on the A-version and the B-version of each of the 21 sets of paired problems. We took two approaches.

First, we conducted a straightforward classical analysis using mean, standard deviation, *t* tests to get a general idea of how the change of representation of the problem situation

Table 2. Number of participants in school types and age groups.

School Type	Age	*n*
Primary (grade 5 and 6)	11–12	969
Secondary	12–18	29,067
Vocational	16–20	1,556

Note: *n* is number of participants

affected each set of paired problems to better understand how the changes in the separate couples of paired problems contributed to the overall result we found (Hoogland, 2016). In this approach each couple of paired problems can be seen as a separate trial, for which in each version around 16,000 participants provided results. By the design of the instrument both groups that provided results for each version can be considered comparable, so a *t* test is a suitable test for a first impression on how the separate couples of paired problems behaved under the change in representation.

Second, we used a probit model which gave the opportunity to investigate more in depth how the difference between the scores on the A and the B versions were interdependent with the variables we measured from the participants (gender, ethnicity, school level, grade level) or with the variables we constructed for this study, that is, wordiness and number of pictures, and content domain of the paired problems. This statistical analysis gives indications in terms of which variables are interdependent with the observed changes in students' results. However, this analysis gives no explanation of what exactly causes the change in students' results. Our main aim was to establish that changing the representation of the problem situation in contextual mathematical problem-solving has indeed an effect on students' results, and that conjectures based on this awareness should be further researched with much more tasks and with various groups of participants. We elaborate on this in the Discussion section.

The probit model and the independent variables

The results on the tasks were scored with wrong/right (0/1) assuming that a right answer occurred when the performance of the student crossed a certain threshold of ability. The independent variables, measured or not, were assumed to have a normal distribution. From these assumptions a probit model – a limited dependent variable model which allows for a multivariate analysis – is the most suitable statistical model to model and analyse the data and present the results. For analysing the overall results in an earlier study this probit model was already composed (Hoogland, 2016) and for this study it is extended with the new defined variables on complexity and domain of the tasks. In this approach each couple of paired problems can be seen as a problem with 31,842 observations. The representation (A-version or B-version) can be regarded as the manipulated variable v with coefficient α_1 which could have an effect on the participants' results. The question was then, if the independent variable v (version of problem) was a significant factor in interpreting the score on the problem. In summary the probit model built up as follows: if ability $y = \alpha_0 + \alpha_1 v + \alpha_2 x + \varepsilon$ crosses a threshold δ then the problem is answered correctly ($z = 1$). We can then formulate the probabilities:

$$P(z = 1) = P(y \geq \delta) = P(\varepsilon \geq \delta - \alpha_0 + \alpha_1 v + \alpha_2 x)$$

and:

$$P(z = 0) = 1 - P(y \geq \delta) = 1 - P(\varepsilon \geq \delta - \alpha_0 + \alpha_1 v + \alpha_2 x)$$

The unknown coefficients $\delta - \alpha_0$, α_1 and α_2 were estimated by maximum likelihood. In all our analyses these coefficients were calculated using STATA11 (probit, dprobit). The background variables we used were school level ($\alpha_{2,1}$ - $\alpha_{2,7}$), grade ($\alpha_{2,8}$), gender ($\alpha_{2,9}$), ethnicity ($\alpha_{2,10}$), age deviation ($\alpha_{2,11}$), and overall mathematics performance ($\alpha_{2,12}$). The

task related variables wordiness ($\alpha_{2,13}$), number of pictures ($\alpha_{2,14}$), and task domain of the task ($\alpha_{2,15}$ - $\alpha_{2,17}$), we used only in this study. All these variables, as non-manipulated independent factors, might contribute to the measured results.

By this design, the characteristics of the participants answering the A-version of a particular task and the characteristics of the participants answering the B-version of that task had the largest likelihood of being the same. This should hold for the measured characteristics as well as for the characteristics that were not measured. The independent manipulated variable was the version of the problem (A or B). The dependent variable was the students' scores on the A-versions and the B-versions of the problems. More details of the probit model and the results of the overall conclusions were described extensively in an earlier publication (Hoogland et al., 2016).

Results

In earlier reports we showed that the overall effect was a significantly higher score of about two percentage points on image-rich numeracy problems, albeit with a very small effect size of Cohen's $d = .09$. For this study we first zoomed in for each paired problem on the results of the t test on the difference in score on both versions. Second, we interpreted the results of probit model to show the effects of the other not-manipulated independent relevant variables. The probit model also provided the opportunity to investigate in depth, if additional variables related to task characteristics could shed further light on the effect that these variables had on the difference in students' performance on both versions of the paired problems.

The results on separate paired problems

We analysed the data by comparing the average scores of the participants on each paired problem, relating the average score on the A-version of a problem with the average score on the B-version of the problem. A two-sided t test with pooled variances was conducted (see Table 3) to evaluate whether for each task the scores on the two versions differed significantly.

To avoid the risk of family-wise errors in our presentation of the results we take into account only tasks that give significant results with $p < .001$ (***). Under this restriction we found eight paired problems (5, 10, 11, 13, 18, 19, 20, 21) where the scores on the B-version were significantly higher than the scores on the A-versions. Furthermore, we found four paired problems (2, 3, 12, 17) for which the scores on the A-versions were significantly higher than the scores on the B-versions.

These results were in line with the small effect we found in the overall results. This gives no cause for straightforward inferences. We see that tasks which "benefits" (B > A) from the change in representing the problem situation are occurring more often than vice versa. In the remainder of the section we analyse whether this can be attributed to specific characteristics of the tasks.

The results of the probit-model

The results of the basic probit model with the background variables school level ($\alpha_{2,1}$ - $\alpha_{2,7}$), grade ($\alpha_{2,8}$), gender ($\alpha_{2,9}$), ethnicity ($\alpha_{2,10}$), age deviation ($\alpha_{2,11}$), and overall

Table 3. The difference in students' performance on task level.

Item	N version A	N version B	M (SE) version A	M (SE) version B	*t* test $P(\|T\|>\|t\|)$
1	15,878	15,964	.716 (.004)	.720 (.004)	.424
2	15,986	15,856	.525 (.004)	.483 (.004)	.000***
3	15,785	16,057	.314 (.004)	.290 (.004)	.000***
4	15,835	16,007	.826 (.003)	.833 (.003)	.131
5	16,038	15,804	.720 (.004)	.828 (.003)	.000***
6	15,775	16,067	.631 (.004)	.640 (.004)	.102
7	16,065	15,777	.404 (.004)	.416 (.004)	.042
8	16,298	15,544	.303 (.004)	.299 (.004)	.420
9	16,069	15,773	.221 (.003)	.213 (.003)	.085
10	15,882	15,960	.495 (.004)	.525 (.004)	.000***
11	15,850	15,992	.145 (.003)	.310 (.004)	.000***
12	15,871	15,971	.466 (.004)	.438 (.004)	.000***
13	15,931	15,911	.619 (.004)	.641 (.004)	.000***
14	15,889	15,953	.040 (.002)	.046 (.002)	.080
15	15,793	16,049	.394 (.004)	.388 (.004)	.264
16	15,921	15,921	.803 (.003)	.815 (.003)	.005
17	15,986	15,856	.803 (.003)	.787 (.003)	.000***
18	15,847	15,995	.153 (.003)	.168 (.003)	.000***
19	15,932	15,910	.247 (.003)	.284 (.004)	.000***
20	15,925	15,917	.130 (.003)	.164 (.003)	.000***
21	16,044	15,798	.188 (.003)	.256 (.003)	.000***

Note: N is number of items tested. M is mean score on items (with standard error in parentheses) $P(|T|>|t|)$ is result of *t* test, unpaired, unequal with hypothesis that difference in score is 0; ***$p < .001$. Version A is the word problem; Version B is the image rich numeracy problem; Significant results are underlined.

mathematics performance ($\alpha_{2,12}$) have been reported earlier (Hoogland, 2016). In Table 4 the analysis is expanded with the variables for wordiness ($\alpha_{2,13}$) and number of depictive elements ($\alpha_{2,14}$).

In Table 4 we present for each variable the calculated coefficients and the marginal effect. This is the (virtual) effect on the students' results of a change of the variable with unit 1 while all the other variables are fixed on their mean value. The marginal effect

Table 4. Probit model expanded with task characteristics: wordiness and number of depictive elements.

Variable		Model expanded with task characteristics regarding wordiness and number of depictive elements. Coefficient(SE)	Marginal effect(SE)
Version (A = 0, B = 1)	α_1	.059*** (.010)	.023*** (.004)
Primary education	$\alpha_{2,1}$	−.547*** (.017)	−.200*** (.005)
Secondary general	$\alpha_{2,2}$ - $\alpha_{2,6}$	−.059*** (013)	−.024*** (.005)
Vocational education	$\alpha_{2,7}$	.328*** (.016)	.130*** (.006)
Grade	$\alpha_{2,8}$	.232*** (.002)	.092*** (.001)
Gender	$\alpha_{2,9}$	.114*** (.003)	.045*** (.001)
Ethnicity	$\alpha_{2,10}$	−.150*** (004)	−.059*** (.002)
Relative age	$\alpha_{2,11}$	−118*** (002)	−.047*** (.001)
Last math grade	$\alpha_{2,12}$	.052*** (001)	.021*** (000)
# Words (Version A)	$\alpha_{2,13}$	−.016*** (.000)	−.006*** (000)
# Depictive elements (Version B)	$\alpha_{2,14}$	−.221*** (.003)	−.087*** (.001)
Constant	α_0*	−.547*** (.017)	

Note: $\alpha_{i,j}$ is the coefficient in the probit model, α_0*= -δ + α_0. Coefficients are calculated with maximum likelihood by STATA11, standard errors are in parentheses. The measure of good fit pseudo-$R^2 = .040$ (McFadden's R^2). All variables are significant ***$p < .001$. The effect of Number of words is only analysed for the word problems (version A); The effect of Number of depictive elements is only analysed for the image-rich numeracy problems (version B).

gives an indication of the size of the effect of that variable. For instance, the overall conclusion of the generic increase of two percentage points as a result of changing the representation of the problem situation (from descriptive to depictive) can be seen in the first row of Table 4, as the marginal effect of the variable *version* going from 0 to 1.

The probit-model expanded with variables wordiness and number of pictorial elements

Two variables were used to indicate the complexity of the problems: wordiness ($\alpha_{2,13}$) and number of pictorial elements ($\alpha_{2,14}$). Table 1 shows for each task the values of these variables. We expanded the probit model with these data and analysed the effect of these variables on the students' performance. Table 4 shows that the wordiness of the word problem had a significant relation to the students' performance, and that the effect was negative, as was expected: higher wordiness led to lower performance on the A-versions of the problem. From Table 4 we also concluded that the number of pictorial elements was significant for the students' performance and that the effect was negative: a higher number of pictorial elements led to lower performance on the B-versions of the problem. This indicated also that the number of pictorial elements seems to be a valid measure of complexity of image-rich numeracy problems. So, by reducing words and adding pictures there could be an optimal combination regarding the effect on students' results. Representing a problem situation with few words and few pictures could be such optimal combination. In this case we limit our inference to the direction of the change and not the size. Follow-up research with larger numbers of items and much greater variability of the variables *number of words* and *number of pictures* could give a better indication of the direction and perhaps the size of such change.

The probit-model expanded with the variable content domain

In Table 5 the results can be found of expanding the probit model with the variables regarding the content domains of the tasks. The used content domains from the Dutch LaNF, numbers ($\alpha_{2,15}$), proportions ($\alpha_{2,16}$), and measurement & geometry ($\alpha_{2,17}$), were represented with seven tasks each (see Table 1).

In Table 5 in the probit analysis the domain of numbers was used as reference category for the effect of the domain categories. The marginal effects regarding the variables proportions and measurement & geometry were negative compared with the variable numbers. This meant that in this instrument the tasks from the domains of proportions and measurement & geometry were harder than the tasks from the domain of numbers for the participating students, which is consistent with what we found in Table 2.

For the aim of this study, however, we looked at the interdependent effects of the variables regarding the content domains and the variable *version*, and we found that the cross term *measurement & geometry* × *version* was significant, and the cross term *proportion* × *version* was not. This meant that the effect of changing the representation of the problem situation had some effect in the domain of measurement & geometry. The effect on tasks in the domain of measurement & geometry was in line with our expectations: the depictive representation in these kinds of tasks was likely to be beneficial for having success in the problem-solving process.

Table 5. Probit model expanded with variables regarding content domain.

		Model expanded with item characteristics regarding content domain	
Variable		Coefficient(SE)	Marginal effect(SE)
Version (A = 0, B = 1)	α_1	.037*** (.006)	.014*** (.002)
Primary education	$\alpha_{2,1}$	−.561*** (.017)	−.205*** (.005)
Secondary general	$\alpha_{2,2}$ - $\alpha_{2,6}$	−.060*** (.013)	−.024*** (.005)
Vocational education	$\alpha_{2,7}$	.335*** (.016)	.133*** (.006)
Grade	$\alpha_{2,8}$	.236*** (.002)	.093*** (.001)
Gender	$\alpha_{2,9}$	.117*** (.003)	.046*** (001)
Ethnicity	$\alpha_{2,10}$	−.153*** (004)	−.060*** (002)
Relative age	$\alpha_{2,11}$	−.120*** (002)	−.047*** (.001)
Last math grade	$\alpha_{2,12}$	.053*** (001)	.021*** (.000)
Numbers	$\alpha_{2,15}$	reference	reference
Proportions	$\alpha_{2,16}$	−.494*** (.006)	−.191*** (.002)
Meas. & Geom.	$\alpha_{2,17}$	−.567*** (.006)	−.218*** (.002)
Numbers × Version	cross term	reference	reference
Proportions × Version	cross term	−.016 (.008)	−.006 (.003)
Meas.& Geom.× Version	cross term	.060*** (.008)	.024*** (.003)
Constant	α_0*	−.675*** (.016)	

Note: $\alpha_{i,j}$ is the coefficient in the probit model, α_0*= -δ + α_0. Coefficients are calculated with maximum likelihood by STATA11, standard errors are in parentheses. The measure of good fit pseudo-R^2 = .056 (McFadden's R^2). All variables are significant ***$p < .001$, with the exception of proportions × version.

Discussion

This study was part of a larger project to research alternatives to word problems in representing the problem situation in mathematical contextual problems. In earlier analyses of the data collected with the described instrument we found as an overall effect that changing the representation of the problem situation from descriptive to mainly depictive had a small positive effect of around 2 percentage point in students' performance. Further analysis showed that the measured effect was not dependent on background variables such as age of students, ability of students, ethnicity and (type of) school, although there was a small interdependency with gender.

In the overall findings we were able to discern some patterns in the differences in performance we found. These patterns revealed that task characteristics could have an effect, which was interdependent with the effect of changing the representation of the problem situation. Those task characteristics could also be used to give indications of possible explanations of the overall results. We focused on specific task characteristics and the interdependence of each task characteristic with the effect of changing the representation of the problem situation, indicated by a difference in students' performance on word problems and image-rich numeracy problems. From the analysis reported in this study we found that the effect of changing representation of the problem situation on the students' results is interdependent with the following task characteristics: (1) the number of words used in the descriptive representation and the number of pictorial elements used in the depictive representation: and (2) the content domain of the task.

Let us elaborate on each of these points. From the literature on use of language in contextual mathematical problems we got an indication that wordiness of the representation of the problem could have a negative effect on the performance of students, and therefore a positive effect on changing the representation of the problem situation from descriptive to depictive (Boonen et al., 2014; Fuchs et al., 2015). This indication was corroborated by our study. Likewise, from the literature on the use of pictorial elements in classroom problems

we got an indication that the number of pictorial elements could have a negative effect on the performance of students (Leppink et al., 2014). This indication was also corroborated by our study. However, the specific effect of changing from words to pictures as the only manipulated variable, such as in this study, is under-researched, with sparse exceptions (Lowrie, Diezmann, & Logan, 2012).

In the Theoretical Background section we argued that Cognitive Load Theory could provide some tentative explanations for the observed differences in students' results. From this theory an increasing number of pictures could have negative effects on students' results due to element interactivity (van Gog, Paas, & Sweller, 2010). This is corroborated by the results from our study. More advanced interpretations from cognitive load theory (Paas, Van Gog, & Sweller, 2010) are worth discussing. In our study we make a shift to more multimodal representations. According to this theory multimodal representations can have a negative effect on students' results, when the depictive elements are considered as extraneous cognitive load. However, multimodal representations can also have a positive effect on students' results, when the multimodality makes it easier for students to mathematise the problem at hand and to engage prior knowledge and experiences. In this interpretation the intrinsic load is reduced by adding depictive elements, resulting in better student results. This interpretation is corroborated by our results, but will need more research on students' thinking when solving such problems with items that are specifically designed to measure cognitive load problems that involve mathematical thinking.

Regarding the domain of the task, we found hardly any research on the interdependent effect of the domain of the task and the representation of the problem situation. One of the findings of our study was that the effect of changing the representation of the problem situation from descriptive to depictive was least strong with tasks from the domains of numbers and proportions, but clearly stronger for the domain of measurement and geometry. From our experience in mathematics education we could argue that in the domain of measurement and geometry the effect would be the largest: the most likely explanation being that pictorial elements in a geometrical or measurement contexts support students in making the mental model of the situation, and as a consequence support them in formulating the mathematical problem correctly, which is one of the crucial steps in the problem-solving cycle (OECD, 2013).

Or otherwise stated, as exemplified in Figure 2: in the domain of measurement & geometry the information can "be read off more directly from the representation" (Schnotz et al., 2010), which means that the inferential step from representation of the problem situation to the mathematical problem is smaller. For tasks in the domain of measurement & geometry word problems can in many cases be considered as quite an inadequate representation of the problem situation: the designer of the tasks describes a geometrical situation in words, which must be "undescribed" and interpreted by the students to make a geometrical mental model and formulate the mathematical problem. Especially in cases where many words are necessary to describe the geometrical situation, the student is likely to be hampered in showing his geometrical problem-solving knowledge and skills. Figure 2 shows an example of a task in the domain of measurement and geometry with a significant higher score on the depictive representation of the problem situation.

We have discussed the above results bearing in mind various limitations. Our study focused on the interdependency of variables with the observed differences in students'

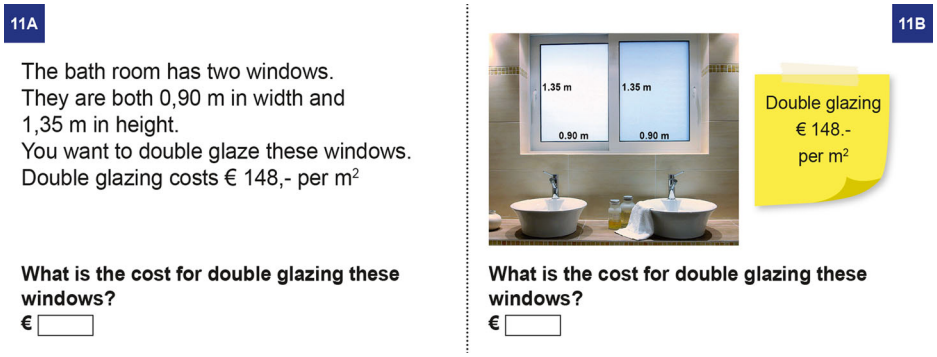

Figure 2. Paired problem 11: Double-glazing.

results. Our study was not designed to single out specific explanations of these differences. Inferences on possible explanations should always consider the following limitations:

First, the research design was based on the assumption that both versions were equivalent in the content, level of mathematical knowledge and skills needed to solve the problem. Moreover, we assumed that the two versions only differed in the representation of the problem situation. In terms of the validity of the measuring tool, various efforts were made to counter threats to validity (Hoogland et al., 2016). The evaluation of the equivalences of the two versions was carried out by experts in the field of mathematics and numeracy education. In their evaluation they were specifically asked to disregard the representation of the problem situation. However, the research literature claims that experts are not necessarily the best evaluators in terms of establishing equivalence or level of assessment items (Baird, Cresswell, & Newton, 2000). Moreover, the intertwined cognitive processes of interpreting, mathematising, and problem-solving are so complex that results involving these processes must be presented with great prudence. Repeated measurements with the same and with other items would be necessary (as a minimum), to see whether and how the results hold up under replication (e.g. Hoogland and Pepin (2017)). More sophisticated research designs could shed more light on the intricacies involved in interpreting verbal and depictive representations of reality, such as research that investigates the ways students actually reason during the problem-solving process, for instance by analysing student work, analysing thinking-aloud protocols or eye-tracking investigations.

Second, in this *post hoc* analysis we used two ways of operationalising task complexity. We chose "wordiness" as a proxy for the complexity of descriptive representations, and the number of pictures as a proxy for the complexity of depictive representations. We found that they had significant effect on the student results and on the change in representation. We acknowledge that these are very crude measures of complexity, which do not account for the effects of more mixed representations. Moreover, we based our analysis on only 21 items. Research on much larger numbers of items would be necessary to see whether the effects and the conclusions are firm enough to take into consideration by designers of assessment tasks. More sophisticated research designs could help to counter these threats to the validity of the findings.

Third, this study did not focus on investigating the actual behaviour of students in solving contextual mathematical problems. Such investigations could shed more light

on the ways students make sense of the posed problems and the interdependence of this with the representation of the problem situation.

Fourth, nowadays students are trained in word problems and are familiar with them. The effect of this on their performance on rather "new" image-rich problems is under-researched yet. It needs more investigations to determine the precise effect of complexity and representation of the problem on performance and the problem-solving mind-set of the students. Changing the mindsets of students to engage more in a problem-solving attitude cannot be reached by merely changing the representation of the problem situation to more realistic and authentic. For a larger effect in such direction a change in classroom behaviour of teachers and a change in dominant assessments practices would be necessary. This study gave some indications how a change in representation of the problem situation in contextual mathematical problems could support such a change.

Conclusion

In educational settings where contextual mathematical problems are used, the effect on students' performance of changing the verbal representation of the problem situation to a more depictive representation of the problem situation is still under-researched. This study is one of very few studies that investigated the effect on students' performance of changing the representation of the problem situation in contextual mathematical problems in a randomised controlled trial. In many countries large-scale assessments have had, and often still have a large influence on decision-making concerning educational policies and curriculum development. In these cases, an increase of a few percentage points in results is significant for decision-makers.

In this study it has not been our expectation to find an overall explanation for the measured effects on performance of students of changing the representation of the problem situation in contextual mathematical problems. There are likely to be more factors than task characteristics such as wordiness and domain of the task that could offer possible explanations of the measured results. More quantitative and qualitative research is needed to get a better understanding of the relation between problem representation and student behaviour, for instance research that investigates the actual reasoning by students when solving problems by thinking-aloud protocols or by stimulated recall.

Professional test designers might benefit from the findings of our study, and they may want to further validate the ways in which they can translate goals of numeracy frameworks into actual test items. Seen in that light it is important to consider that changing the representation of the problem situation to a more authentic and image-rich form is likely to have an effect on student results.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Ministerie van Onderwijs, Cultuur en Wetenschap (Dutch Ministry of Education) under the programme Onderwijsbewijs II (Evidence-based Research in Education), under grant number ODB10068.

ORCID

Kees Hoogland http://orcid.org/0000-0002-0650-6999

References

Baird, J.-A., Cresswell, M., & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, *15*(2), 213–229. doi:10.1080/026715200402506

Boonen, A. J. H., van Wesel, F., Jolles, J., & van der Schoot, M. (2014). The role of visual representation type, spatial ability, and reading comprehension in word problem solving: An item-level analysis in elementary school children. *International Journal of Educational Research*, *68*(0), 15–26. doi:http://doi.org/10.1016/j.ijer.2014.08.001

CBS. (2012). *Jaarboek Onderwijs in cijfers* 2011. Retrieved from Heerlen.

Cito instituut voor toetsontwikkeling. (2015). Voorbeeldtoets 2F 2014 [Example numeracy test level 2F 2014]. Retrieved from http://www.cito.nl/~/media/cito_nl/files/voortgezet%20onderwijs/cito_voorbeeldtoets_2f_2014.ashx

Daro, P. (Producer). (2013). Phil Daro - Against "Answer Getting". [Video] Retrieved from https://vimeo.com/79916037

Depaepe, F., De Corte, E., & Verschaffel, L. (2010). Teachers' approaches towards word problem solving: Elaborating or restricting the problem context. *Teaching and Teacher Education*, *26*(2), 152–160. doi:http://doi.org/10.1016/j.tate.2009.03.016

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

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

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

Geiger, V., Goos, M., & Forgasz, H. (2015). A rich interpretation of numeracy for the 21st century: A survey of the state of the field. *ZDM*, *47*(4), 531–548. doi:10.1007/s11858-015-0708-1

Gill, G. T., & Hicks, R. C. (2006). Task complexity and informing science: A synthesis. *Informing Science Journal*, *9*, 1–30.

Gravemeijer, K. (1997). Solving word problems: A case of modelling? *Learning and Instruction*, *7*(4), 389–397. doi:10.1016/s0959-4752(97)00011-x

Greer, B. (1997). Modelling reality in mathematics classrooms: The case of word problems. *Learning and Instruction*, *7*(4), 293–307. doi:http://doi.org/10.1016/S0959-4752(97)00006-6

Hoogland, K. (2016). *Images of numeracy: Investigating effects of visual representations of problem situations in contextual mathematical problem solving*. (PhD-thesis), Technical University Eindhoven, Eindhoven, The Netherlands.

Hoogland, K., & De Koning, J. (2013). *Dataset: Rekenen in beeld [Dataset: Images of numeracy]*. http://doi.org/10.17026/dans-za6-5q6c

Hoogland, K., & Pepin, B. (2017). The intricacies of assessing numeracy: Investigating alternatives to word problems. *Adults Learning Mathematics - An International Journal*, *11*(2), 14–26.

Hoogland, K., Pepin, B., Bakker, A., de Koning, J., & Gravemeijer, K. (2016). Representing contextual mathematical problems in descriptive or depictive form: Design of an instrument and validation of its uses. *Studies in Educational Evaluation*, *50*, 22–32. doi:10.1016/j.stueduc.2016.06.005

Hoogland, K., & Stelwagen, R. (2011). *A new Dutch numeracy framework*. Paper presented at the Adults Learning Mathematics , 18th International Conference, Dublin, Ireland.

Leppink, J., Paas, F., van Gog, T., van der Vleuten, C. P. M., & van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, *30*(0), 32–42. doi:http://doi.org/10.1016/j.learninstruc.2013.12.001

Lowrie, T., Diezmann, C., & Logan, T. (2012). A framework for mathematics graphical tasks: The influence of the graphic element on student sense making. *Mathematics Education Research Journal*, *24*(2), 169–187. doi:10.1007/s13394-012-0036-5

Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *Timss 2015 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

OECD. (2013). *PISA 2015 draft Mathematics Framework*. Retrieved from Paris, France: http://www.oecd.org/pisa/pisaproducts/Draft20PISA20201520Mathematics20Framework%20.pdf

Paas, F., Van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review*, *22*(2), 115–121. doi:10.1007/s10648-010-9133-8

Palm, T. (2009). Theory of authentic task situations. In L. Verschaffel, B. Greer, W. V. Dooren, & S. Mukhopadhyay (Eds.), *Words and worlds - modelling verbal descriptions of situations* (pp. 3–20). Rotterdam, the Netherlands: Sense.

PIAAC Numeracy Expert Group. (2009). *PIAAC Numeracy: A conceptual framework*. Retrieved from Paris, France: /content/book/9789264128859-en http://dx.doi.org/10.1787/9789264128859-en

Piel, S., & Schuchart, C. (2014). Social origin and success in answering mathematical word problems: The role of everyday knowledge. *International Journal of Educational Research*, *66*(0), 22–34. doi:http://doi.org/10.1016/j.ijer.2014.02.003

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27–57. doi:10.1093/applin/22.1.27

Schnotz, W., Baadte, C., Müller, A., & Rasch, R. (2010). Creative thinking and problem solving with depictive and descriptive representations. In L. Verschaffel, E. d. Corte, T. d. Jong, & J. Elen (Eds.), *Use of representations in reasoning and problem solving - analysis and improvement* (pp. 11–35). London, UK: Routledge.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–370). New York, NY: McMillan.

Schoenfeld, A. H. (2014). Reflections on learning and cognition. *ZDM*, *46*(3), 497–503. doi:10.1007/s11858-014-0589-8

Thompson, A. G., Philipp, R. A., Thompson, P. W., & Boyd, B. A. (1994). Calculational and conceptual orientations in teaching mathematics. In A. Coxford (Ed.), *1994 yearbook of the NCTM* (pp. 79–92). Reston, VA: NCTM.

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

Verschaffel, L., Corte, E. D., & Lasure, S. (1994). Realistic considerations in mathematical modeling of school arithmetic word problems. *Learning and Instruction*, *4*(4), 273–294. doi:10.1016/0959-4752(94)90002-7

Verschaffel, L., Depaepe, F., & Van Dooren, W. (2014). Word problems in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 641–645). Dordrecht, the Netherlands: Springer.

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

Verschaffel, L., Greer, B., Van Dooren, W., & Mukhopadhyay, S. (Eds.) (2009). *Words and worlds - modelling verbal descriptions of situations*. Rotterdam, the Netherlands: Sense.

Watson, A., & Ohtani, M. (Eds.) (2015). *Task design in mathematics education - an ICMI study 22*. Cham, Switzerland: Springer International.

Williams, G., & Clarke, D. J. (1997). The complexity of mathematics tasks. In N. Scott & H. Hollingsworth (Eds.), *Mathematics: Creating the future* (pp. 451–457). Melbourne, Australia: AAMT.

Yackel, E., & Cobb, P. (1995). Classroom sociomathematical norms and intellectual autonomy. In L. Meira & D. Carraher (Eds.), *Proceedings of the nineteenth international conference for the psychology of mathematics education* (Vol. 3, pp. 264–271). Recife, Brazil: Program Committee of the 19th PME conference.