Studies in Educational Evaluation 58 (2018) 122–131

Contents lists available at ScienceDirect

Studies in Educational Evaluation

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

Changing representation in contextual mathematical problems from descriptive to depictive: The effect on students' performance

Kees Hoogland[a,*], Jaap de Koning[b], Arthur Bakker[c], Birgit E.U. Pepin[d], Koeno Gravemeijer[d]

[a] HU University of Applied Sciences Utrecht, Utrecht, The Netherlands
[b] SEOR, Erasmus University Rotterdam, Rotterdam, The Netherlands
[c] Freudenthal Institute, Utrecht University, Utrecht, The Netherlands
[d] Eindhoven School of Education, Eindhoven University, Eindhoven, The Netherlands

ARTICLE INFO

Keywords:
Contextual mathematical problems
Numeracy
Descriptive
Depictive
Controlled randomised trial

ABSTRACT

Research on solving mathematical word problems suggests that students may perform better on problems with a close to real-life representation of the problem situation than on word problems. In this study we pursued real-life representation by a mainly depictive representation of the problem situation, mostly by photographs. The prediction that students perform better on problems with a depictive representation of the problem situation than on comparable word problems was tested in a randomised controlled trial with 31,842 students, aged 10–20 years, from primary and secondary education. The conclusion was that students scored significantly higher on problems with a depictive representation of the problem situation, but with a very small effect size of Cohen's $d = 0.09$. The results of this research are likely to be relevant for evaluations of mathematics education where word problems are used to evaluate the mathematical capacity of students.

1. Introduction

In mathematics education worldwide it is common practice (Verschaffel, Greer, & De Corte, 2000; Verschaffel, Greer, Van Dooren, & Mukhopadhyay, 2009) to use word problems to teach and assess students in the domain of solving quantitative problems from real life. 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" (Verschaffel et al., 2000, p. ix)

Word problems can be seen as a special genre of contextual mathematics problems. Many studies and discourses, however, give rise to serious concerns as to whether word problems foster or accurately assess students' potential to solve quantitative problems from everyday life (Gellert & Jablonka, 2009; Gerofsky, 2009, 2010; Gravemeijer, 1997; Greer, 1997; Roth, 2009). Students very often do not succeed in bridging the gap between their school mathematical knowledge and representations of real-life situations in word problems (Boaler, 1993). Studies show that students in a classroom setting persistently approach word problems with a calculational orientation (Thompson, Philipp, Thompson, & Boyd, 1994) and typically do not take into account realistic considerations about the problem situation (Cooper & Harries, 2002, 2003; Dewolf, Van Dooren, Ev Cimen, & Verschaffel, 2014; Dewolf, Van Dooren, Hermens, & Verschaffel, 2015; Verschaffel, Corte, & Lasure, 1994). This behaviour leads to underachievement; the students do not make sense of the problem situation, and hence cannot show their full potential in solving problems from daily life. There are indications from several studies that this effect might be counteracted or avoided by making the problem situations or the representation of the problem situations more authentic for the students (Palm, 2002, 2006, 2008, 2009; Verschaffel et al., 2009).

For the study reported here we designed an alternative to existing word problems whereby a descriptive representation of the problem situation, as is common in word problems, was replaced as much as possible with a depictive representation of the problem situation that we assumed to be closer to real life. In the design process we paid special attention to the real-life familiarity and relevance of the chosen images. We called the resulting problems "image-rich numeracy problems". In the spirit of the aforementioned definition of word problems, we suggested the following definition for image-rich numeracy problems: Image-rich numeracy problems can be defined as visual representations of a problem situation wherein one or more questions are raised, the answer to which can be obtained by the application of mathematical reasoning to numerical data available in the problem representation.

* Corresponding author.
E-mail address: keeshoogland1@outlook.com (K. Hoogland).

https://doi.org/10.1016/j.stueduc.2018.06.004
Received 28 November 2017; Received in revised form 11 April 2018; Accepted 13 June 2018

In our study we compared students' performance on word problems with their performance on image-rich numeracy problems, which were mathematically equivalent to the word problems and in which images were used to make the problem situation closer to students' real-life experiences. We designed an instrument to compare students' performance; this consisted of a web-based test with 21 paired problems. The instrument was used in a randomised controlled trial with over 32.000 Dutch students.

The primary aim of this study was to establish whether the performance of students changed due to changing the representation of the problem situation. Given the difficulties that have been reported with word problems, we predicted that the performances could be better, because some of the aforementioned difficulties with word problems and the resulting underachievement could be counteracted with the change of representation. The ubiquitous use of word problems in mathematical assessment all over the world makes this research relevant for practitioners and test developers from many countries. The next section elaborates on the difficulties that were reported in using word problems in schools and how we explored ways to counteract them by changing the representation of the problem situation.

2. Theoretical background

2.1. Difficulties with word problems

Over the last 20 years the use of word problems in assessing mathematical potential in solving real-life problems has been intensively researched. Many research findings mention serious drawbacks of using straightforward word problems to assess students' mathematical potential (Verschaffel et al., 2000, 2009). Numerous studies on student behaviour when solving word problems report various student blockages in the problem-solving process (Galbraith & Stillman, 2006). The most reported behaviour is that students base their analysis and calculations on a rather arbitrary association between certain salient quantitative elements of the problem situation and certain mathematical operations (Thompson et al., 1994; Verschaffel et al., 2000). Studies of word problems further show that students tend not to consider the possible constraints imposed by reality (Caldwell, 1995; Cooper & Harries, 2003; Dewolf et al., 2015; Lave, 1992; Reusser & Stebler, 1997; Schoenfeld, 1992; Verschaffel et al., 1994, 2009; Wyndhamn & Säljö, 1997). Students seem to value the bare outcome of the calculation more than the realism of the outcome. There is strong evidence that students approach word problems as "school math" problems and not as problems from real life. This behaviour of students is reported to be reinforced by teachers' approaches that tend to be more on the mathematical structure of the problem than on the contextual aspects (Depaepe, De Corte, & Verschaffel, 2010).

An explanation for this behaviour is the strong calculational orientation (Thompson et al., 1994) that for many decades has been, and to a great extent still is (Madison & Steen, 2008), dominant in mathematics classrooms. Thompson et al. (1994) distinguish between a calculational approach in which the focus is primarily on procedures and operations with numbers, and a conceptual approach in which the focus is primarily on explaining, reasoning, and interpreting the quantitative situation. In the calculational approach, quantitative problems seem first and foremost to be used to train students in executing arithmetical operations. From this perspective, the representation of the problem situation is not an especially important aspect and the word problems are mainly straightforward. When teachers and students share a calculational orientation, they maintain a mathematics classroom culture based on implicit or explicit sociomathematical norms, in which solving word problems is limited to finding a calculation to perform without relating the outcomes to the original problems. (Gravemeijer, 1997; Yackel & Cobb, 1995).

The use of superficial strategies and dissociation from reality in problem-solving with word problems has become known as "suspension of sense-making" (Reusser & Stebler, 1997; Schoenfeld, 1991; Verschaffel et al., 2000). This phenomenon is quite persistent in classroom situations and has generated serious concern as to whether students can show their full potential in solving quantitative real-life problems when confronted with word problems in a classroom or in an assessment setting.

2.2. Counteracting difficulties with word problems

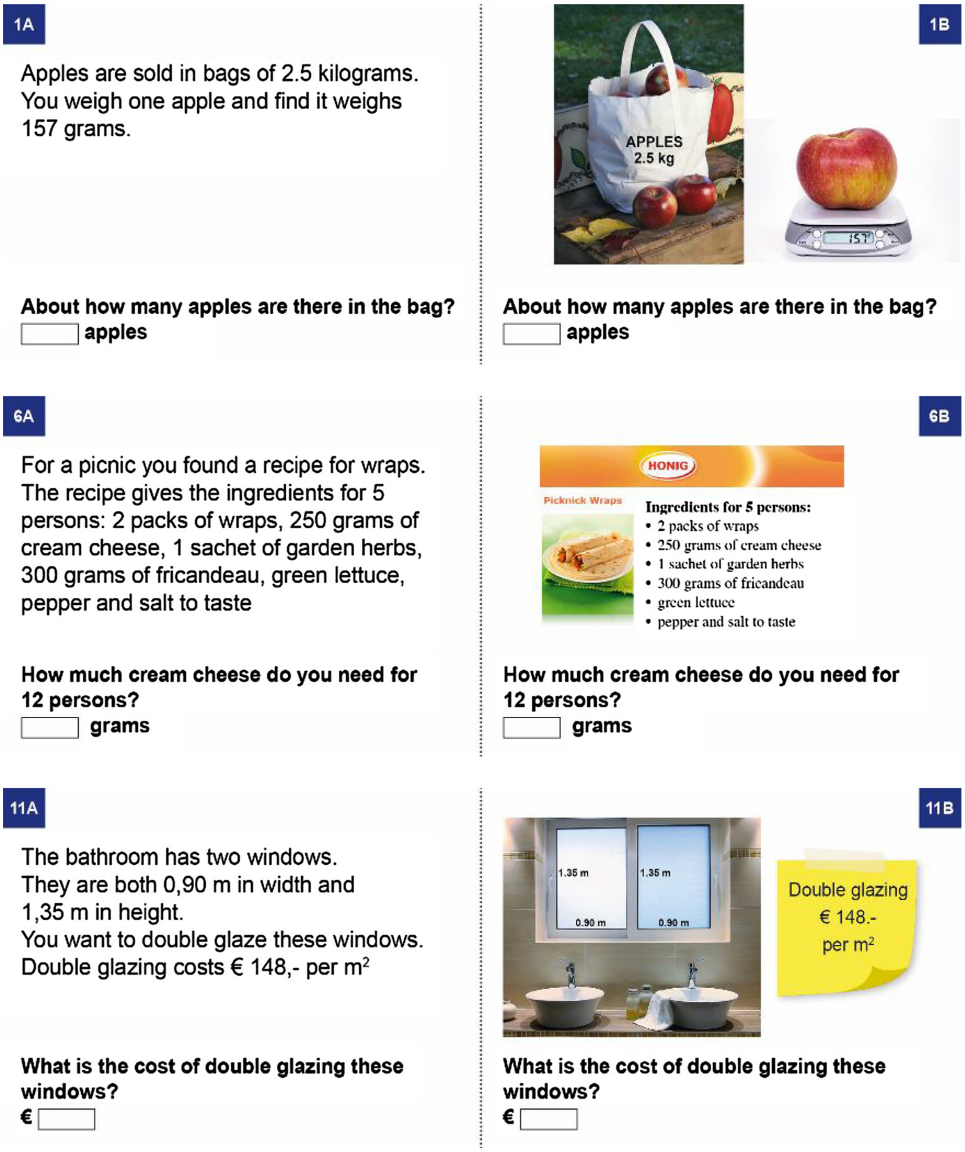
Many attempts have been made to counteract or avoid the calculational approach and the suspension of sense-making by reinforcing a more conceptual approach by the students, for instance, by adapting word problems into less straightforward problems; by adding instructions for students to take realistic considerations; and by changing the setting in which problem-solving takes place (Cooper & Harries, 2002, 2003; DeFranco & Curcio, 1997; Dewolf, Van Dooren, & Verschaffel, 2011; Palm, 2009; Reusser & Stebler, 1997; Verschaffel et al., 2000; Wyndhamn & Säljö, 1997). To counteract suspension of sense-making, most of these studies suggest making the problem situation or the representation of the problem situation more authentic for students, hence helping them to make sense of it and focus more on a conceptual approach.

Other researchers and practitioners question the feasibility of counteracting suspension of sense-making by adapting word problems. They advocate the creation of real-life situations in mathematics lessons in order to teach and assess students' potential to deal with quantitative problems from everyday life (Bonotto, 2007, 2009; Frankenstein, 2009; Lave, 1992; Zevenbergen & Zevenbergen, 2009). Bonotto (2007) recommends encouraging students to analyse mathematical facts embedded in appropriate "cultural artefacts", such as supermarket receipts, bottle and can labels, railway schedules, or a weekly TV guide. Although many of the arguments for using real-life situations to teach students relevant problem-solving skills are convincing, there is no widespread dissemination of such practices. Practical constraints and the persistent classroom culture of "getting to the right answer as quickly as possible" are mentioned as the main reasons (Verschaffel et al., 2000). There is strong evidence that students are taught to approach word problems as "school maths" problems and not as problems of authentic life situations. This student behaviour is said to be reinforced by teachers' approaches that tend to foreground the mathematical structure of the problem rather than the contextual aspects (Depaepe et al., 2010).

2.3. Towards a more effective design of depictive representations

In our particular attempt to design problems using images from real-life situations to avoid suspension of sense-making we took into account other studies that investigated the effect of using pictorial elements in contextual mathematical problems. Some of these studies may indicate a mitigating effect on the decrease of suspension of sense-making that we predicted.

In the early stages of educational psychology there was a focus on individual characteristics. The Dual Coding Theory (Paivio, 1986), for instance, categorised people as visual learners (visualisers) or verbal learners (verbalisers). Massa and Mayer (2006), however, found no empirical evidence that in order to gain better results verbal learners should be given verbal instruction and visual learners should be given visual instruction. Instead they found that adding pictorial aids to an online lesson that was heavily text-based tended to help both visualisers and verbalisers. An earlier study by Plass, Chun, Mayer, and Leutner (1998) on learning a second language also concluded that a combination of text and pictures yielded better results in terms of learning outcomes than text alone. Research findings from this theoretical perspective add to the plausibility of our prediction. Mayer (2005) defined multimedia learning as learning from words (e.g., spoken or printed text) and pictures (e.g., illustrations, photos, maps, graphs, animation



1A

Apples are sold in bags of 2.5 kilograms.
You weigh one apple and find it weighs
157 grams.

About how many apples are there in the bag?
□ **apples**

1B

About how many apples are there in the bag?
□ **apples**

6A

For a picnic you found a recipe for wraps.
The recipe gives the ingredients for 5
persons: 2 packs of wraps, 250 grams of
cream cheese, 1 sachet of garden herbs,
300 grams of fricandeau, green lettuce,
pepper and salt to taste

**How much cream cheese do you need for
12 persons?**
□ **grams**

6B

Ingredients for 5 persons:
- 2 packs of wraps
- 250 grams of cream cheese
- 1 sachet of garden herbs
- 300 grams of fricandeau
- green lettuce
- pepper and salt to taste

**How much cream cheese do you need for
12 persons?**
□ **grams**

11A

The bathroom has two windows.
They are both 0,90 m in width and
1,35 m in height.
You want to double glaze these windows.
Double glazing costs € 148,- per m^2

**What is the cost of double glazing these
windows?**
€ □

11B

Double glazing
€ 148.-
per m^2

**What is the cost of double glazing these
windows?**
€ □

Fig. 1. Three examples of paired problems: Word Problem (WP) and Image-Rich Problem (IRP). Our transslation.

or video). Our image-rich numeracy problems fall into this definition because, next to the representation of the problem situation, the problem question is posed in words. Various studies from the perspective of multimedia learning suggest that, under the right conditions, the combination of text and pictures can promote better comprehension (Mayer, 2005).

In other studies on using more depictive representations in (word) problem solving, lower scores (Berends & van Lieshout, 2009; Kaminski, Sloutsky, & Heckler, 2008) or no effect (Dewolf et al., 2014, 2015) have been reported. These studies differed from our study in the sense that images in our study were authentic situations, relevant and an integral part of the problem situation. In our design of image-rich numeracy problems we tried to keep as close as possible to the actual problems to be solved, to keep students as much as possible away from suspension of sense-making.

We have also considered research investigating specifically the solving of (word) problems by students, for instance, students of mixed ability, and their use of pictorial elements in the problem-solving process (Boonen, 2015; Hegarty, 2004; Krawec, 2014; Van Garderen & Montague, 2003; Van Garderen, 2006). The results in these studies corroborate our chain of reasoning. These studies showed that when students in the solving process use drawings that are structured and relevant to the mathematical model needed, the performance of students improves. However, when they make drawings that are merely pictorial or not relevant, their performance is worse, as is to be expected.

Research on using text and pictures in lesson material not always points in the same direction. In recent years, research on performance in mathematical problems has been carried out from a Cognitive Load Theory perspective (Sweller, 2005, 2010). These studies show that redundancy of information, common when adding illustrations, can put extra cognitive load on students' working memories, and split-attention effects can occur when students have to jump between text and illustration elements (Berends & van Lieshout, 2009; Rasmussen & Bisanz, 2005; Scheiter, Gerjets, & Catrambone, 2006). From this perspective, an opposite effect could occur on the student results, contrary to the positive effect we predicted on students' performance from using more real-life elements in the representation of problem situations.

An overall perspective is provided by research on the effect of depictive and descriptive representations on creativity and problem-solving (Schnotz & Bannert, 2003; Schnotz, 2002; Schnotz, Baadte, Müller, & Rasch, 2010). Schnotz and Bannert (2003) concluded that task-appropriate graphics may support learning and task-inappropriate graphics may interfere with mental model construction. Schnotz et al.

(2010) stated that to solve a quantitative problem, a task-oriented construction of a mental mathematical representation is necessary, provided that it is task-appropriate. Schnotz's line of reasoning is that depictive representations can help students make a relevant mathematical mental model of the situation, and that depictive representations have a high inferential power because the information can "be read off more directly from the representation" (p. 21). This perspective also added to the plausibility of our prediction.

2.4. The images used

In our study we selected as set of word problems that were used in a variety of high-stakes mathematics assessments of 10–20 years-old students in the Netherlands spread over several mathematical domains. As an alternative to each word problem we designed problems that could be used in regular assessment situations, and at the same time would be close to real-life problem situations, by changing the description of the problem situation to a mainly depictive representation. We replaced the verbal – sometimes verbose – representation of the problem situation with a representation that is visually connected to the quantitative problem at hand. With technologies such as digital cameras and on-screen presentations, this has become a feasible option in regular teaching and assessment situations. We followed the reasoning in Palm (2009) which states: "that a strong argument can be made that the fidelity of the simulations (…) clearly has an impact on the extent to which students, when dealing with school tasks, may engage in the mathematical activities attributed to the real situation that are simulated" (p. 9).

In our depictive representations, the problem situation was represented with images from real-life situations, in the form of photographs, headlines from newspapers and "handwritten" notes (see Fig. 1), and in that way augmenting the fidelity of the simulation, as advocated by Palm (2009). We distinguished our research from research on word problem solving that uses or adds images that are less realistic or sometimes possibly confusing for students (Berends & van Lieshout, 2009; Dewolf et al., 2014, 2015).

Considering the evidence from the abovementioned research, we expected that the use of images depicting real situations would increase the students' association with real-life situations and problems and therefore would decrease suspension of sense-making and possibly the strong calculational orientation. In the trial, this was translated to the hypothesis that students scored better on our image-rich numeracy problems than on comparable word problems.

2.5. The hypothesis

In this study we predicted that students would score better on image-rich numeracy problems than on comparable (only) word problems. Theories and research from educational psychology indicate that it matters how authentic and relevant the used images are. Therefore, in our research project we designed image-rich numeracy problems that were as close as possible to the realistic problem they represented. The pairs of problems we designed, are available under open access (Hoogland & De Koning, 2013). Subsequently, we put our prediction to the test. Assessing students' potential to solve quantitative problems from daily life is a multifaceted problem in which many factors play a role. In this study one variable, namely type of representation of the problem situation, was systematically varied. In our analysis we also took into account the interaction effects between the manipulated variable – descriptive versus depictive version – and background factors of the participants, such as school type, grade level, gender, ethnicity, and age. The number of collected data allowed for such an in-depth analysis, by which the possible interacting effects of these background variables could be ruled out. Considering the literature, we expected a positive but small effect. From our power analysis, therefore, we aimed at a number of participants well over 5000, so that inferences about a small effect could be drawn (Ellis, 2010). This is elaborated on in the method section.

3. Method

For the international reader we provide information on the context in which the instrument was designed and trialled. In the Netherlands the relevance of the developed instrument is high, as in 2010 a "Referentiekader Taal en Rekenen" [Literacy and Numeracy Framework (LaNF)] was passed as law (Ministerie van OCW, 2009). In this framework four content domains were formulated: numbers; proportions; measurement & geometry; and relations (tables, diagrams, graphs, formulas, etc.). The instrument we designed can be used by schools as a diagnostic tool for upcoming nationwide examinations on the content of the LANF.

3.1. Design

The design and validation of the instrument to measure the effect of changing the representation of the problem situation form descriptive to depictive is described extensively in an earlier article (Hoogland, Pepin, Bakker, de Koning, & Gravemeijer, 2016). We summarised the main elements of this design and validation in the next paragraphs.

The instrument consisted of a web-based numeracy test, in which 21 problems came in either one of two versions: word problem or image-rich numeracy problem. The instrument called for problems in two versions: a word problem version (WP) and an image-rich numeracy problem version (IRP) (e.g. Fig. 1).

The design of the paired problems was conducted as follows. First, 40 word problems were selected from existing (Dutch) high-stakes numeracy tests, and second, 40 image-rich items were designed with the same problem situation and the same problem question. Hence, a batch of 40 paired problems in the mathematical domains of the LANF were constructed. A panel of eight independent experts in mathematics and numeracy education was asked to analyse the 40 paired problems and answer the following questions: (1)" Do the two versions of the problem test the same mathematical knowledge and skills?", (2) "If the two versions test the same mathematical knowledge and skills, are they testing on the same mathematical level?". This led to 21 paired problems for which the first question was answered positively by all, or all but one, of the experts. The experts' answers to the second question helped us for the final instrument, so that the selected problems were spread evenly around the domains of the LaNF. The original 40 paired problems and the final instrument with 21 paired problems in both Dutch and English are available under open access (Hoogland & De Koning, 2013). In an earlier article we reported on measures to counter threats to validity (Hoogland et al., 2016).

To each participant, a test was delivered of 21 problems presented in either the IRP version or the WP version, with a minimum of 9 and a maximum of 12 of each version, randomly assigned. The order of the problems in the test situation was again randomised for each participant. By this design the trial fulfilled the conditions of a randomised controlled experiment – the characteristics of the participants answering the IRP version of a particular problem and the characteristics of the participants answering the WP version of that problem are highly likely to be the same. This holds for the measured characteristics as well as for the characteristics that were not measured. The manipulated variable is the *version* of the problem (WP or IRP); the dependent variable is the student's score on the problems.

3.2. Participants

In a four-week period, the test was available on the internet for schools to use as a numeracy test. In total, 31,842 participants, from 179 schools geographically spread across the Netherlands, took the test. Table 1 shows the number of participants from different age groups and

Table 1
Number of Participants in School Types and Age Groups.

School Type	Subtype	Age	*n*
Primary (BO)		11–12	969
Pre-vocational (VMBO)	Low (VMBO-BB)	12–16	1,932
	Middle (VMBO-KB)	12–16	2,658
	High (VMBO-GT)	12–16	7,869
General secondary (HAVO)		12–17	8,918
Pre-university (VWO)		12–18	7,670
Vocational (MBO)		16–20	1,146
Unknown			680

Note. n is number of participants.

different educational streams in the Dutch school system. From the total student population in the Netherlands around 2% participated.

In the Dutch school system, primary education is for 4- to 12-year olds and runs over eight grades (K-6). In secondary education, the Netherlands has a highly streamed school system. VMBO is a (pre-) vocational education stream for 12- to 16-year olds; HAVO and VWO are the general secondary and pre-university streams that prepare students for college and university respectively. MBO is a tertiary vocational stream that is a follow-up to VMBO and is intended for 16- to 20-year olds.

Of the participants, 49.5% were male, 49.1% were female and for the remainder the gender was unknown. Of the participants, 23% were considered to be of a migrant family and 77% not. In line with the definition by Statistics Netherlands (CBS) a participant was considered to be of a migrant family if the participant or one of his/her parents was born outside the Netherlands (CBS, 2012). The percentages of gender and ethnicity in the various school types were close to the national percentages (CBS, 2012). Therefore, we assumed participating schools to be representative of Dutch schools in general.

3.3. Tasks

Each problem in the web-based test was presented as screen-filling. The question was posed at the bottom of the screen. Below the question, the numerical solution to the problem could be entered and was computer-scored as right (1) or wrong (0). Fig. 1 provides three examples of paired problems with a WP version and an IRP version.

3.4. Procedure

Each participant was assigned a personal activation code to begin the digital test with the 21 problems. Participants conducted the test on-screen at an internet-connected PC. For the total test, a time limit of 60 min was set, which was sufficient for the participants to complete the test. An online calculator was allowed in solving the problems. All answers to the problems were numerical values, to be entered into an answer field by the participants.

The participants' answers for each problem were scored and recorded. After finishing the test, a short digital questionnaire was administered to each participant to collect the following additional data: school type, grade level, gender, ethnicity, age and math grade (test or student report) last received. All data were recorded anonymously in a research database.

After the experiment the test as a whole was evaluated by analysing the mean score for each item across the whole cohort and the item-rest-correlations. The correlation between the score on one item and the average score of the other items can reveal whether any item in the set is (in)consistent with the averaged behaviour of the others, and if inconsistent, should be discarded. The analysis showed that none of the items failed to fulfil the requirements of an acceptable item. The mean scores of the items range from 0.04 to 0.84 and all items had positive item-rest correlations between 0.05 and 0.54. The Cronbach α for the test was 0.84, which indicated good internal consistency (Kline, 1999).

3.5. Statistical analysis

To test the prediction, we first carried out a paired-samples *t*-test on the mean scores of the students on their IRP version items and their WP version items. To get more in-depth insight into the effects of the background variables on the results and their possible interaction effect with the manipulated variable, we analysed the data through a model approach. We choose for the most fitting model approach for this kind of data, namely a probit model, which is a limited dependent variable model (Long, 1997). In a probit model the following assumptions are common: the probability that a student solves a given problem is determined by his or her "proven ability" of crossing a certain threshold. This proven ability may differ from true ability because students may not be able to show their true ability fully (Borooah, 2002). Our prediction suggested that this is the case with word problems more than with image-rich numeracy problems, because students' abilities were impaired by characteristics of word problems that hinder the demonstration of their full potential.

We can then build the model of proven ability $y = \alpha_0 + \alpha_1 v + \alpha_2 x + \varepsilon$ and formulate the probabilities of answering an item correctly:

$$P(z = 1) = P(y \geq \delta) = P(\varepsilon \geq \delta - \alpha_0 - \alpha_1 \mathrm{v} - \alpha_2 \mathrm{x}) \text{ and}$$

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

Proven ability *y* was not observed directly, but was only indirectly reflected in the student's results in solving the presented problems. We assumed that a problem is answered correctly if proven ability *y* crosses a threshold value δ. The dichotomous variable *z* took the value 1 when the problem was answered correctly ($y \geq \delta$) and 0 when it was answered incorrectly. We further assumed that proven ability *y* was normally distributed and depended on several background variables that as non-manipulated independent factors contributed to the measured outcomes on the problems: school type, grade level, gender, ethnicity, age and mathematics grade last received. In the model $y = \alpha_0 + \alpha_1 v + \alpha_2 x + \varepsilon$, *v* is the manipulated dummy variable *version*: for word problems *version* = 0; for image-rich problems *version* = 1. The vector *x* can consist of one or more of the non-manipulated independent variables presented in Table 2. The error term ε represented unobserved variables affecting ability. In the model ε can be treated as a random variable with a normal distribution with mean 0 and y is scaled in such a way that the variance of ε is equal to 1. So, y defined a probit model

Table 2
Overview of Recorded Non-manipulated Independent Variables.

Variable		Explanation	Coding
BO	$\alpha_{2,1}$	primary education	dummy
VMBO-BB	$\alpha_{2,2}$	pre-vocational, lower level	dummy
VMBO-KB	$\alpha_{2,3}$	pre-vocational, intermediate level	dummy
VMBO-GT	$\alpha_{2,4}$	pre-vocational, higher level	dummy
HAVO	$\alpha_{2,5}$	secondary general	dummy
VWO	$\alpha_{2,6}$	pre-university education	dummy
MBO	$\alpha_{2,7}$	secondary vocational	dummy
Grade	$\alpha_{2,8}$		BO(5–6), VMBO(1–4), HAVO (1–5), VWO(1–6), MBO(1–4)
Gender	$\alpha_{2,9}$		0 = female; 1 = male;
Ethnicity	$\alpha_{2,10}$		0 = not migrant family; 1 = migrant family;
Agedev	$\alpha_{2,11}$	age relative to average age in grade level	0–3
Maths grade	$\alpha_{2,12}$	maths grade last received	1–10

Note. $\alpha_{i,}$j is the corresponding coefficient in the probit model with reversed sign. For all dummy variables the coding is: 0 = not in this level, 1 = in this level.

from which consistent estimators for the coefficients α_1 and α_2 can be obtained by maximizing the likelihood function.

The participants were asked for their age, but since age is dependent on grade level, we changed the variable age to relative age (Agedev) as the deviation of a participant's age to the average age of all participants in the same grade level. Maths grade was the last test grade or report mark the participant received for his or her mathematics performance. It was an indication of the mathematical ability of a participant relative to the abilities of other participants in the same school level and grade level.

4. Results

We first conducted a classical analysis of the test results to get an overall idea of the average students score on the word problems and on the image-rich numeracy problems. Second, we conducted an in-depth analysis using a probit model, taking into account the possible effects of other variables and the possible interdependency of the involved variables.

4.1. Correct scores of the participants

We analysed the data as a set of 31,842 paired observations, where for each participant a pair of observations consisted of the mean score of that participant on the WP version items and the mean score of that participant on the IRP version items presented in his or her test. We conducted a paired samples *t*-test on the mean correct scores on items of both versions and found M(A) = .436 (.234), M(B) = .455 (.237), and a difference of .019(.202) with $t = 16.84$, which was statistically significant ($p < .001$), and we found as effect size Cohen's $d = 0.09$. This result indicated that the students' average performance on the IRP version items was a statistically significant two percentage points higher than on the WP version items. The data therefore supported our prediction. However, the effect size of Cohen's $d = 0.09$ could be considered very small.

In Table 3 we present the results of correct scores on all separate items and we distinguish for the main conditions: school type, gender, and ethnicity. For all these subgroups we found the same pattern.

4.2. Results from the probit analysis of three expanding models

The probit model was used to investigate how the background variables, alongside the manipulated variable (IRP version versus WP version), contributed to the correct scores on the items. In the probit model we considered the data as a set of 31,842 times 21, which is 668,682 item scores. In this approach we considered the trial as 668,682 observations of student behaviour on individual items. The probit model analysis related the probability that a student solves an item correctly to a set of independent variables.

Table 3
Correct Scores of Participants on word problems (A-version) and inage-ricgh problems (B-version).

		A-version *M* (SD)	B-version *M* (SD)	*n*
All items		0.44(0.44)	0.45(0.45)	31,842
	BO	0.24 (0.42)	0.25(0.43)	969
	VMBO_BB	0.18(0.38)	0.19(0.39)	1,932
	VMBO_KB	0.26(0.44)	0.28(0.45)	2,658
	VMBO_GT	0.36(0.48)	0.37(0.48)	7,869
	HAVO	0.49(0.50)	0.50(0.50)	8,918
	VWO	0.60(0.49)	0.62(0.48)	7,670
Condition	MBO	0.52(0.50)	0.54(0.50)	1,146
	Gender = f	0.42(0.49)	0.44(0.50)	15,637
	Gender = m	0.46(0.50)	0.48(0.50)	15,766
	Ethnicity = not migrant family	0.45(0.50)	0.47(0.50)	24,183
	Ethnicity = migrant family	0.39(0.49)	0.41(0.44)	7,220

Note. *n* is number of participants. *M* is mean score over all participants and per condition with standard deviation in parenthesis.

In Table 4 we present three runs of the probit analysis with an expanding number of non-manipulated independent variables. Given the large number of participants and the small differences in scores, we used p < .001 for significance. In model 1, VMBO-BB, VMBO-KB, VMBO-GT, HAVO, VWO were treated as an aggregated group and acted as the reference group.

From Table 4 we concluded that the positive effect of the IRP version on a correct score on an item is significant and stayed significant if other variables were taken into account. As a caveat note that the values of the coefficients are not indicative of the magnitude of the effect, but the sign of the coefficients give its direction. Table 4 showed that adding variables hardly has any effect on the coefficients. This meant that the model and the outcomes are quite robust and only depended to a small extent on a particular choice of variables.

Table 5 presents the marginal effects calculated from model 3. In the probit model the marginal effect of a variable is the increase in probability of answering an item correctly under the condition that all other variables are at their mean. The marginal effects can be considered as a measure of effect size, but this must be done with great prudence (Hoetker, 2007).

The marginal effects presented in Table 5 can be interpreted as follows. The variable *version* had two values: 0 for the word problem version and 1 for the image-rich version. Changing the representation of the problem from the WP version to the IRP version resulted in an increase of two percentage points in overall students' performance. Let us compare this with the effect of gender. The variable gender has also two values: 0 for female and 1 for male. The conclusion is that male students scored overall five percentage points higher than female students on the problems presented.

From the marginal effects we concluded that when the problem is presented with the IRP version rather than with the WP version, the chance of a fictitious participant answering an item correctly increased by around two percentage points, assuming that all other variables are at their mean. This was consistent with what we found in Table 3.

The variables regarding school type (*BO, VMBO-BB, VMBO-KB, VMBO-GT, HAVO, VWO, MBO*) were statistically significant and the sign of the coefficient was consistent with the increasing levels. The variable *gender* was statistically significant and the coefficient indicated that males had a higher chance than females of getting an item correct. The variable *ethnicity* was statistically significant and the coefficient indicated that the chance of answering an item correctly was lower for participants from migrant families than for participants from non-migrant families. The variable *math grade* was statistically significant and the coefficient indicated that participants with a higher math grade had a higher chance of answering an item correctly.

Overall we concluded that the variable *version* was significant, but with a very small effect. This effect was comparable to the effect of being from a non-migrant family or of having a higher math grade. The effect was less than that of gender. In the next section we analysed in more depth the interaction effects of these variables.

4.3. In-depth analysis with interaction terms

An in-depth analysis was carried out by investigating how the manipulated variable *version* interacted with the other, non-manipulated, variables. In our model the variable *version* v is the manipulated variable: for word problems $v = 0$ and for image-rich problems $v = 1$. The interaction between variables can be examined by including so-called interaction terms in the probit model. For instance, the interaction between the variable *version* and the variable *BO* can be examined by incorporating a variable *version * BO* in the model – this indicates the degree of interaction. In the first column of Table 6 seven interaction

Table 4
Coefficients of Manipulated and Non-manipulated Variables in Three Models.

Variable		**Coefficients** **model 1**	**Coefficients** **model 2**	**Coefficients** **model 3**
Version	α_1	0.05*(0.00)	0.05*(0.00)	0.05* (0.00)
BO	$\alpha_{2,1}$	-0.40*(0.01)	-0.52*(0.01)	-0.60*(0.01)
VMBO-BB	$\alpha_{2,2}$		-0.92*(0.01)	-0.95*(0.01)
VMBO-KB	$\alpha_{2,3}$		-0.62*(0.01)	-0.64*(0.01)
VMBO-GT	$\alpha_{2,4}$		-0.33*(0.00)	-0.34*(0.00)
HAVO	$\alpha_{2,5}$	ref. cat.	ref. cat.	ref. cat.
VWO	$\alpha_{2,6}$		0.29*(0.00)	0.26*(0.00)
MBO	$\alpha_{2,7}$	0.35*(0.01)	0.23*(0.01)	0.26*(0.01)
Grade	$\alpha_{2,8}$	0.22*(0.00)	0.22*(0.00)	0.24*(0.00)
Gender	$\alpha_{2,9}$	.		0.13*(0.00)
Ethnicity	$\alpha_{2,10}$			-0.04*(0.00)
Agedev	$\alpha_{2,11}$			-0.01*(0.00)
Maths grade	$\alpha_{2,12}$			0.06* (0.00)
Unknown var.	α_0^*	-0.66*(0.00)	-0.55*(0.01)	-1.02* (0.01)
pseudo-R^2		.02	.07	.08

Note. αi,j is the coefficient in the probit model, $\alpha_0^* = \delta - \alpha_0$. Coefficients are calculated with maximum likelihood by STATA11, standard errors are in parentheses. Ref.cat is reference category. The measure of good fit pseudo-R^2 is McFadden's R^2. All variables are significant, *$p < .001$.

Table 5
Marginal Effects of Manipulated and Non-manipulated Variables in Probit Model of Participants Correct Scores.

Variable		Marginal effects (SE)
Version	α_1	0.02 *(0.00)
BO	$\alpha_{2,1}$	−0.22* (0.00)
VMBO-BB	$\alpha_{2,2}$	−0.32* (0.00)
VMBO-KB	$\alpha_{2,3}$	−0.23* (0.00)
VMBO-GT	$\alpha_{2,4}$	−0.13* (0.00)
HAVO	$\alpha_{2,5}$	ref. cat.
VWO	$\alpha_{2,6}$	0.10* (0.00)
MBO	$\alpha_{2,7}$	0.10* (0.00)
Grade	$\alpha_{2,8}$	0.09* (0.00)
Gender	$\alpha_{2,9}$	0.05* (0.00)
Ethnicity	$\alpha_{2,10}$	−0.02* (0.00)
Agedev	$\alpha_{2,11}$	−0.00*(0.00)
Math grade	$\alpha_{2,12}$	0.02* (0.00)

Note. Marginal effects are calculated with dprobit in STATA11, standard errors are in parentheses. $\alpha_{i,}$j is the coefficient in the probit model. Ref. cat. is reference category. All variables are significant *$p < .001$.

terms were added.

After running the model, Table 6 showed that none of the interaction terms were significant. Although interpreting interaction terms in non-linear models is hazardous (Ai & Norton, 2003; Greene, 2010), the global picture showed that the effect of the variable *version* was independent of school type.

Next, we added in the probit model the interaction terms of the variable *version* with the variables *gender*, *ethnicity*, *age deviation*, and *math grade last received* respectively. In the first column of Table 7 these five interaction terms can be seen. After running the model, Table 7 showed that none of these interaction terms were significant.

Furthermore, we checked the interaction between being in a specific school and the variable *version* to investigate whether better scores on

Table 6
Probit Model Coefficients and Marginal Effects of Manipulated and Non-manipulated Variables on School Type.

Variable	Coefficient (SE)	Marginal effect (SE)
Version	0.05 (0.01)	0 .02 (0.00)
BO	−0.52*(0.01)	−0.19* (0.00)
VMBO-BB	−0.93*(0.01)	−0.31* (0.00)
VMBO-KB	−0.64*(0.01)	−0.23* (0.00)
VMBO-GT	−0.33*(0.01)	−0.13* (0.00)
HAVO	ref. cat.	ref. cat.
VWO	0.28*(0.01)	0.11* (0.00)
MBO	0.23*(0.01)	0.09* (0.00)
Grade level	0.22*(0.00)	0.09* (0.00)
BO ⁎ Version	−0.01 (0.02)	0.00 (0.01)
VMBO-BB ⁎ Version	0.02 (0.02)	0.01 (0.01)
VMBO-KB ⁎ Version	0.03 (0.01)	0.01 (0.01)
VMBO-GT ⁎ Version	0.00 (0.01)	0.00 (0.00)
HAVO ⁎ Version	ref. cat.	ref. cat.
VWO ⁎ Version	0.01 (0.01)	0.00 (0.00)
MBO ⁎ Version	0.01 (0.02)	0.00 (0.01)
Unknown Variables	−0.54* (0.01)	

Note. Coefficients (and standard errors) and marginal effects (and standard errors) are displayed. Ref. cat. is reference category. Variables are significant, interaction terms are not significant with *$p < .001$.

Table 7
Probit Model Coefficients and Marginal Effects of Manipulated and Non-manipulated Variables on Participants' Background.

Variable	Coefficient	Marginal effect
Version	0.05** (0.02)	0.02** (0.01)
Gender	0.10* (0.00)	0.04* (0.00)
Ethnicity	−0.14* (0.01)	−0.05* (0.00)
Agedev	−0.12* (0.00)	−0.05* (0.00)
Maths grade	0.02* (0.00)	0.01* (0.00)
Gender * Version	0.02 (0.01)	0.01 (0.00)
Ethnicity * Version	−0.00 (0.01)	−0.00 (0.00)
Agedev * Version	0.01 (0.00)	0 .00 (0.00)
Maths grade * Version	−0.00 (0.00)	−0.00 (0.00)
Unknown Variables	−0.33 (0.01)	

Note. Coefficients (and standard errors) and marginal effects (and standard errors) are displayed. Variables are significant, interactions are not significant with *p < .001.

IRP versions could be explained by the fact that some schools specially trained students for these kinds of problems. In the probit model we checked for this interaction by incorporating each of the 179 schools as a separate variable and studying the interaction terms with the variable *version*. The interaction terms *school nnn * version* were not significant. There was no indication that the effect of the variable *version* was caused by specific schools.

5. Discussion and conclusion

Students' difficulties with the genre of word problems is a serious educational challenge. These difficulties are becoming increasingly relevant since word problems are used more and more for high stakes testing and international comparison studies (Hoogland & Stelwagen, 2011; OECD, 2012; Palm & Burman, 2004; PIAAC Numeracy Expert Group, 2009). Several researchers (Gravemeijer, 1997, 2009; Greer, 1997; Palm, 2008) have theorised on improvement of word problems in assessment situations and the way they are used in classroom practice. In addition, there is ample empirical research on the effects on students' performance of changing the design of the problems or of changing the classroom setting in which the problem solving takes place (DeFranco & Curcio, 1997; Reusser & Stebler, 1997; Wyndhamn & Säljö, 1997). A common factor in the findings is that making the problems more real-life could reduce both the "suspension of sense making" (Schoenfeld, 1991) and the predominant calculational approach (Thompson et al., 1994), and could result in better student performance.

The present study extended that body of knowledge by investigating the effects of changing specifically one aspect of existing word problems – the way in which the problem situation is represented – from primarily descriptive or verbal, as is common in word problems, to primarily depictive, using photographs as representations of real-life situations. This is commonly used in contemporary multimedia lesson materials and multimedia assessment tools (Mayer, 2005). In our study we paid special attention to the authenticity and the relevance of the depictive elements. As the goal is to assess whether students can apply their mathematical concepts in real-life situations, the use of representations of authentic situations seems self-evident. At the same time, one can argue that every representation of "reality" creates its own reality or its own genre of realities (Gerofsky, 2009, 2010). Nevertheless, there are indications that more depictive representation of the problem situation can counteract to a certain extent the difficulties students have with word problems. From a cognitive psychology perspective it is argued (Schnotz et al., 2010) that a more depictive representation can help students make a relevant (mental) mathematical model of the situation, which can lead to more success in solving the problem.

Comparing our results with those of other similar studies, it has become clear that positive effects on students' results can only be established when the used images are close to the real-life problem situation, relevant and an integral part of the problem situation. This could explain why in others studies lower scores have been reported (Berends & van Lieshout, 2009) or no effect (Dewolf et al., 2014, 2015). The results of our study can also be contrasted with results of studies that focus on the use of drawings by students as part of the problem-solving process. Most of those studies found that when students used relevant and structured diagrams they performed better, and when they used more pictorial sketches they performed worse (Boonen, van Wesel, Jolles, & van der Schoot, 2014; Elia, Gagatsis, & Demetriou, 2007; Hegarty, 2004; Krawec, 2014; Van Essen & Hamaker, 1990; Van Garderen, 2006). In all these cases the relevance and appropriateness of the depictive elements were of major importance.

From the perspective of Cognitive Load Theory (Sweller, 2010) one can argue that redundancy and split-attention effects due to using visual representations of the problem situation in a classroom problem, could have a negative effect on better result. Weighing the arguments of the different perspectives, in the present study the prediction was that students would score better on numeracy problems where the problem situation was represented as closer to real life by using relevant and "realistic" depictive elements.

Reviewing the literature, we expected a small positive effect on students' results by replacing the descriptive representation of the problem situation with a more depictive one. This was actually what we found: the effect of a mainly depictive representation on participants' correct scores was around two percentage points, which was statistically significant with a very small effect size. Although the effect was small, the result of a two percentage points increase is a noteworthy effect, in particular in large-scale assessments.

Changing the representation of the problem situation was not a teaching intervention, unlike the teaching programs and long-lasting interventions that were compared, for instance, by Hattie (2009, 2015), who reported an average effect size of .40 for major educational interventions that ranged over several months. It would be surprising or even dubious if the changes we made should have an effect size that could compare with that. The effect size found is comparable with the findings of Slavin (2016) on large randomised studies in education. Because of the expected small effect size, we paid special attention to the power analysis. To be able to draw inferences about such small effect sizes a large sample is necessary. Ellis (2010) suggests a sample of well over 5000 participants to be able to draw inferences about effect sizes around $d = 0.10$. The sample in our study was big enough to fulfil these conditions. The measured rise in performance of 1.9 percentage points in our study came with a 95% confidence interval of [1.7; 2.1]. Combining this with the fact that a two percentage-point rise in performance in large-scale testing could make the difference between passing and failing for many students – in our sample around 600 students –, we were inclined to conclude we had a noteworthy effect, which needed, however, further research and analysis to come to possible recommendations for practical use.

This conclusion was strengthened further by excluding other factors that could cause this effect. Because of the large number of participants, we were able to investigate whether the variable *version* was interacting with background variables of the participants by incorporating interaction terms of the variable *version* with background variables and with specific school variables. We checked this with a significance level of *p < .001. None of the other variables had significant interaction with the variable *version.* We found that the measured result is robust, meaning that the variable *version* was the sole factor that explained the increase in students' results.

In conclusion, we found that the students' better scores on our image-rich numeracy problems than on comparable word problems could be attributed with high certainty to the representation of the problem situation. This result supported our prediction that using visual elements in the representation of the problem situation close to a real-life problem situation could possibly prevent to a certain extent the suspension of sense-making and the calculational approach that occur so frequently when students are solving word problems in classroom and assessment situations.

Further research could investigate whether specific task characteristics had an influence on the measured effect. For instance, the domain of the tasks and the wordiness of the word problems could be of effect. Other future research to find possible explanations for the measured effect could be in the realm of solution times, strategy selection, problem-solving behaviour, or other explanations on the cutting edge of mathematics education and cognitive and neurosciences.

The results of this study should also be interpreted with an eye on their limitations. One limitation was the small number of 21 paired items used in the trial. If we had anticipated that more than 30,000 students would participate, we could have pooled more items, a strategy used in international assessments such as PISA (OECD, 2012) and PIAAC (PIAAC Numeracy Expert Group, 2009). Furthermore, the study focused on students' scores and not on their actual behaviour. The actual behaviour of students when solving image-rich numeracy problems could be quite different from their behaviour in solving word problems. Investigating this behaviour, for example by observation or by eye-tracking, could offer more detailed explanations for the results in this study. It could also give further insights into the balance between the positive effect on the scores from decreasing suspension of sense-making and the possible negative effects on the scores caused by redundancy and split-attention effects.

In our view, our findings can act as a good starting point for further research into the effects of representation of problem situations on students' performances. Furthermore, the results could be an incentive for designers of teaching materials and assessments to (re)consider the chosen representations of reality in their products. Overlooking the body of literature on solving contextual mathematical problems and the role of depictive elements in the design of the problems and/or in the solving of the problems by the students, we suggested a short checklist regarding the images used:

- Are the images realistic and relevant for solving the problem (in contrast to cartoon-like and distracting)?
- Are the images an integral part of the representation of the problem situation (in contrast to mere illustration)?
- Are the images relevant and consistent with the envisioned mathematical concepts and mathematical models that are involved in the problem solving (in contrast to irrelevant or even confusing)?

Using such a checklist could make studies on this topic and their results more understandable and comparable.

Using word problems in assessing mathematics and numeracy tests is not a typical Dutch phenomenon. In many countries around the world verbal descriptions of real-life phenomena are common practice, in textbooks and assessment/test items. Hence, the results of this study is of relevance for anyone involved in using word problems in text books, in tests and in (international) assessments. In large scale evaluations of educational attainment, such as PISA (OECD, 2017) and PIAAC (OECD, 2016), it would be appropriate to develop a heightened awareness that choosing a representation of reality in assessing students' skills could influence students' results.

Author note

The research in this article was conducted with a grant from the Dutch Ministry of Education in the program Onderwijsbewijs II [Evidence-based research in education] with grant number ODB10065.

References

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters, 80*(1), 123–129. http://dx.doi.org/10.1016/s0165-1765(03)00032-6.

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

Boaler, J. (1993). Encouraging the transfer of' School' mathematics to the' Real World' through the integration of process and content, context and culture. *Educational Studies in Mathematics, 25*, 341–373.

Bonotto, C. (2007). How to replace the word problems with activities of realistic mathematical modeling. In W. Blum, P. L. Galbraith, H.-W. Henn, & M. Niss (Eds.). *Modelling and applications in mathematics education – The 14th ICMI study* (pp. 297–314). New York, NY: Springer International.

Bonotto, C. (2009). Working towards teaching realistic mathematical modelling and problem posing in Italian classrooms. In L. Verschaffel, B. Greer, W. V. Dooren, & S. Mukhopadhyay (Eds.). *Words and worlds – Modelling verbal descriptions of situations* (pp. 297–314). Rotterdam, The Netherlands: Sense.

Boonen, A. J. H. (2015). *Comprehend, visualize & calculate - Solving mathematical problems in contemporary math education. (Doctoral dissertation).* Amsterdam, The Netherlands: Vrije Universiteit.

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. http://dx.doi.org/10.1016/j.ijer.2014.08.001.

Borooah, V. K. (2002). *Logit and probit: Ordered and multinomial models.* Thousand Oaks, CA: Sage.

Caldwell, L. (1995). *Contextual considerations in the solution of children's multiplication and division word problems. (MA Master's Thesis).* Belfast, Nothern Ireland: Queen's University Belfast.

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

Cooper, B., & Harries, T. (2002). Children's responses to contrasting `realistic' mathematics problems: Just how realistic are children ready to be? *Educational Studies in Mathematics, 49*(1), 1–23. http://dx.doi.org/10.1023/a:1016013332659.

Cooper, B., & Harries, T. (2003). Children's use of realistic considerations in problem solving: Some English evidence. *The Journal of Mathematical Behavior, 22*(4), 449–463. http://dx.doi.org/10.1016/j.jmathb.2003.09.004.

DeFranco, T. C., & Curcio, F. R. (1997). A division problem with remainder embedded across two contexts: Children's solutions in restrictive versus real world settings. *Focus on Learning Problems in Mathematics, 19*(2), 58–72.

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. http://dx.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. http://dx.doi.org/10.1080/00220973.2012.745468.

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

Dewolf, T., Van Dooren, W., & Verschaffel, L. (2011). Upper elementary school children's understanding and solution of a quantitative problem inside and outside the mathematics class. *Learning and Instruction, 21*(6), 770–780. http://dx.doi.org/10.1016/j.learninstruc.2011.05.003.

Elia, I., Gagatsis, A., & Demetriou, A. (2007). The effects of different modes of representation on the solution of one-step additive problems. *Learning and Instruction, 17*(6), 658–672. http://dx.doi.org/10.1016/j.learninstruc.2007.09.011.

Ellis, P. D. (2010). *The essential guide to effect sizes.* Cambridge, UK: Cambridge University Press.

Frankenstein, M. (2009). Developing a criticalmathematical numeracy through real real-life word problems. In L. Verschaffel, B. Greer, W. V. Dooren, & S. Mukhopadhyay (Eds.). *Words and worlds – Modelling verbal descriptions of situations* (pp. 111–130). Rotterdam, the Netherlands: Sense.

Galbraith, P., & Stillman, G. (2006). A framework for identifying student blockages during transitions in the modelling process. *Zentralblatt fuer Didaktik der Mathematik, 38*(2), 143–162. http://dx.doi.org/10.1007/BF02655886.

Gellert, U., & Jablonka, E. (2009). "I am not talking about reality": Word problems and the intracies of producing legitimate text. In L. Verschaffel, B. Greer, W. V. Dooren, & S. Mukhopadhyay (Eds.). *Words and worlds – Modelling verbal descriptions of situations* (pp. 39–54). Rotterdam, the Netherlands: Sense.

Gerofsky, S. (2009). Genre, simulacra, impossible exchange, and the real: How postmodern theory problematises word problems. In L. Verschaffel, B. Greer, W. V. Dooren, & S. Mukhopadhyay (Eds.). *Words and worlds – Modelling verbal descriptions of situations* (pp. 21–38). Rotterdam, the Netherlands: Sense.

Gerofsky, S. (2010). The impossibility of 'real-life' word problems (according to Bakhtin, Lacan, Zizek and Baudrillard). *Discourse: Studies in the Cultural Politics of Education, 31*(1), 61–73. http://dx.doi.org/10.1080/01596300903465427.

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

Greene, W. (2010). Testing hypotheses about interaction terms in nonlinear models. *Economics Letters, 107*(2), 291–296. http://dx.doi.org/10.1016/j.econlet.2010.02.014.

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

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyes relating to achievement.* Oxfordshire, UK: Routledge.

Hattie, J. (2015). *What doesn't work in education: The politics of distraction.* London, UK: Pearson.

Hegarty, M. (2004). Dynamic visualizations and learning: Getting to the difficult questions. *Learning and Instruction, 14*(3), 343–351. http://dx.doi.org/10.1016/j.learninstruc.2004.06.007.

Hoetker, G. (2007). The use of logit and probit models in strategic management research: Critical issues. *Strategic Management Journal, 28*(4), 331–343. http://dx.doi.org/10.1002/smj.582.

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

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. http://dx.doi.org/10.1016/j.stueduc.2016.06.005.

Hoogland, K., & Stelwagen, R. (2011). A New Dutch numeracy framework. *Paper presented at the Adults Learning Mathematics (ALM), 18th international conference (ALM)*.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. *Science, 320*(5875), 454–455. http://dx.doi.org/10.1126/science.1154659.

Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). London, UK: Routledge.

Krawec, J. L. (2014). Problem representation and mathematical problem solving of students of varying math ability. *Journal of Learning Disabilities, 47*(2), 103–115. http://dx.doi.org/10.1177/0022219412436976.

Lave, J. (1992). Word problems: A microcosm of theories of learning. In P. Light, & G. Butterworth (Eds.). *Context and cognition: Ways of learning and knowing* (pp. 74–92). New York, NY: Harvester Wheatsheaf.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.

Madison, B. L., & Steen, L. A. (2008). *Calculation vs. context: Quantitative literacy and its implications for teacher education* (1st ed.). Washington, DC: Mathematical Association of America.

Massa, L. J., & Mayer, R. E. (2006). Testing the ATI hypothesis: Should multimedia instruction accommodate verbalizer-visualizer cognitive style? *Learning and Individual Differences, 16*(4), 321–335. http://dx.doi.org/10.1016/j.lindif.2006.10.001.

Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning.* New York: Cambridge University Press.

Ministerie van OCW (2009). *Referentiekader taal en rekenen [Literacy and numeracy framework].* Retrieved from: http://www.taalenrekenen.nl/downloads/referentiekader-taal-en-rekenen-referentieniveaus.pdf/.

OECD (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills.* Paris, France: OECD Publishing.

OECD (2016). *Technical report of the survey of adult skills (PIAAC)*(2nd edition). Paris, France: OECD Publishing.

OECD (2017). *PISA 2015 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* Paris, France: OECD Retrieved from: http://www.oecd.org/pisa/pisaproducts/PISA%202012%20framework%20e-book_final.pdf.

Paivio, A. (1986). *Mental representations: A dual coding approach.* New York: Oxford University Press.

Palm, T. (2002). *The realism of mathematical school tasks: Features and consequences.* Umea Universitet.

Palm, T. (2006). Word problems as simulations of real-world situations: A proposed framework. *For the Learning of Mathematics, 26*(1), 42–47. http://dx.doi.org/10.2307/40248523.

Palm, T. (2008). Impact of authenticity on sense making in word problem solving. *Educational Studies in Mathematics, 67*(1), 37–58. http://dx.doi.org/10.1007/s10649-007-9083-3.

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.

Palm, T., & Burman, L. (2004). Reality in mathematics assessment : An analysis of task-reality concordance in Finnish and Swedish national assessment. *Nordic Studies in Mathematics Education, 9*(3).

PIAAC Numeracy Expert Group (2009). *PIAAC numeracy: A conceptual framework.* Paris, France: OECDhttp://dx.doi.org/10.1787/9789264128859-en.

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology, 90*(March (1)), 25–36 1998.

Rasmussen, C., & Bisanz, J. (2005). Representation and working memory in early arithmetic. *Journal of Experimental Child Psychology, 91*(2), 137–157. http://dx.doi.org/10.1016/j.jecp.2005.01.004.

Reusser, K., & Stebler, R. (1997). Every word problem has a solution: The suspension of reality and sense-making in the culture of school mathematics. *Learning and Instruction, 7*(4), 309–327. http://dx.doi.org/10.1016/s0959-4752(97)00014-5.

Roth, W.-M. (2009). On the problematic of word problems – Language and the world we inhabit. In L. Verschaffel, B. Greer, W. V. Dooren, & S. Mukhopadhyay (Eds.). *Words and worlds: Modelling verbal descriptions of situations* (pp. 55–69). Rotterdam, the Netherlands: Sense.

Scheiter, K., Gerjets, P., & Catrambone, R. (2006). Making the abstract concrete: Visualizing mathematical solution procedures. *Computers in Human Behavior, 22*(1), 9–25. http://dx.doi.org/10.1016/j.chb.2005.01.009.

Schnotz, W. (2002). Commentary: Towards an integrated view of learning from text and visual displays. *Educational Psychology Review, 14*(1), 101–120. http://dx.doi.org/10.1023/A:1013136727916.

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction, 13*(2), 141–156. http://dx.doi.org/10.1016/s0959-4752(02)00017-8.

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. (1991). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.). *Informal reasoning and education* (pp. 311–343). Hillsdale, NJ: Lawrence Erlbaum.

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.

Slavin, R. E. (2016). *What is a large effect size?* Retrieved fromhttp://www.huffingtonpost.com/robert-e-slavin/what-is-a-large-effect-si_b_9426372.html.

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.). *The Cambridge handbook of multimedia learning* (pp. 31–48). New York, NY: Cambridge University Press.

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

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 Essen, G., & Hamaker, C. (1990). Using self-generated drawings to solve arithmetic word problems. *Journal of Educational Research, 83*(6), 301–312.

Van Garderen, D. (2006). Spatial visualization, visual imagery, and mathematical problem solving of students with varying abilities. *Journal of Learning Disabilities, 39*(6), 496–506. http://dx.doi.org/10.1177/00222194060390060201.

Van Garderen, D., & Montague, M. (2003). Visual-spatial representation, mathematical problem solving, and students of varying abilities. *Learning Disabilities Research & Practice, 18*(4), 246–254. http://dx.doi.org/10.1111/1540-5826.00079.

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. http://dx.doi.org/10.1016/0959-4752(94)90002-7.

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. (2009). *Words and worlds – Modelling verbal descriptions of situations.* Rotterdam, The Netherlands: Sense.

Wyndhamn, J., & Säljö, R. (1997). Word problems and mathematical reasoning—A study of children's mastery of reference and meaning in textual realities. *Learning and Instruction, 7*(4), 361–382. http://dx.doi.org/10.1016/s0959-4752(97)00009-1.

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

Zevenbergen, R., & Zevenbergen, K. (2009). The numeracies of boatbuilding: New numeracies shaped by workplace technologies. *International Journal of Science and Mathematics Education, 7*(1), 183–206. http://dx.doi.org/10.1007/s10763-007-9104-9.