Studies in Educational Evaluation 50 (2016) 22–32

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Representing contextual mathematical problems in descriptive or depictive form: Design of an instrument and validation of its uses

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

[a] SLO – Netherlands Institute for Curriculum Development, Enschede, The Netherlands
[b] Eindhoven School of Education, Eindhoven University, Eindhoven, The Netherlands
[c] Freudenthal Institute, Utrecht University, Utrecht, The Netherlands
[d] SEOR, Erasmus University Rotterdam, Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 4 November 2015
Received in revised form 4 May 2016
Accepted 16 June 2016
Available online 1 July 2016

Keywords:
Word problem
Contextual mathematical problems
Design
Image-rich numeracy problem
Validation

ABSTRACT

The aim of this study is to contribute to the body of knowledge on the use of contextual mathematical problems. Word problems are a predominant genre in mathematics classrooms in assessing students' ability to solve problems from everyday life. Research on word problems, however, reveals a range of difficulties in their use in mathematics education. In our research we took an alternative approach: we designed image-rich numeracy problems as alternatives for word problems. A set of word problems was modified by systematically replacing the descriptive representation of the problem situation by a more depictive representation and an instrument was designed to measure the effect of this modification on students' performance. The instrument can measure the effect of this alternative approach in a randomized controlled trial. In order to use the instrument at scale, we made this instrument also usable as a diagnostic test for an upcoming nationwide examination on numeracy. In this article we explain and discuss the design of the instrument and the validation of its intended uses.

© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In mathematics education there is an increasing focus on the usability of acquired mathematical knowledge and skills (Kilpatrick, 1996; De Lange, 1999; Toner, 2011), and hence there is a growing need for materials and tools to teach and assess the use of mathematical knowledge and skills in real-life situations. For decades it has been common practice to use word problems to teach and assess students' ability to solve quantitative problems in practical day-to-day situations (Verschaffel, Greer, & De Corte, 2000). However, the current practice of using word problems to assess students' ability to solve quantitative problems from everyday life also gives rise to serious concerns: The question is whether word problems are adequate for this purpose (Verschaffel et al., 2000; Verschaffel, Greer, Van Dooren, & Mukhopadhyay, 2009).

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). For this study we use word problems for which both the description of the problem situation as well as the actual problem statement are presented in words.

The reported difficulties with word problems are so persistent that in this study we investigated an alternative for word problems as means to evaluate students' ability to solve quantitative problems in practical day-to-day situations. In this alternative the descriptive representation of the problem situation, as is common in word problems, is replaced as much as possible by a depictive representation, which means using visual elements, mostly photographs, that were as close as possible to the real-life problem situation. To contrast them with word problems we call these problems image-rich numeracy problems. The choice for a more depictive representation of the problem situation is informed by research on difficulties with word problems, (Verschaffel et al., 2009), considerations on the sometimes problematic relation between language, context and sense-making in solving word problems (Sepeng, 2013), considerations on authenticity in mathematical problem solving (Palm, 2009; Verschaffel et al.,

* Corresponding author at: SLO, Netherlands Institute for Curriculum Development, Piet Heinstraat 12, 7511 JE Enschede, the Netherlands.
E-mail addresses: K.Hoogland@slo.nl (K. Hoogland), B.E.U.Pepin@tue.nl (B. Pepin), A.Bakker4@uu.nl (A. Bakker), deKoning@ese.eur.nl (J. de Koning), Koeno@Gravemeijer.nl (K. Gravemeijer).

http://dx.doi.org/10.1016/j.stueduc.2016.06.005
0191-491X/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

2000), and research on problem solving in cognitive psychology (Schnotz, 2002, 2005; Schnotz, Baadte, Müller, & Rasch, 2010). These research perspectives combined strongly suggests that using real-life images, such as photographs, to represent the problem situation has a better chance of keeping students in a problem solving mindset instead of falling back to an answer-getting mindset (Daro, 2013). Photographs are more easily associated with real-life situations, and arguably can feel more authentic for students, and therefore increase the chance that students continue using considerations of reality in the problem solving activities. Furthermore, it is likely that language and text comprehension difficulties are reduced by more depictive representations of the problem situation. The effects of changing the representation of the problem situation on students' performance are still under-researched and we decided a validated instrument was needed to measure these effects. In this article we described the design of such an instrument and the validation of its intended uses. The instrument was subsequently used in the Dutch context (Hoogland, De Koning, Bakker, Pepin, & Gravemeijer, submitted) and the English version of the instrument is now available under open access (Hoogland & De Koning, 2013).

2. Theoretical perspectives

2.1. Mathematics as usable knowledge

Over the past fifty years situations from real life have increasingly been used in the school mathematics classrooms. There are several developments in mathematics education that have pushed this trend forward. First, there is a plea by mathematics educators for a stronger emphasis in school mathematics on the ways in which mathematics is used in daily life (Kilpatrick, 1996; De Lange, 1999). Kilpatrick (1996) observes that "the curriculum had shifted (. . .) away from an emphasis on abstract structures towards efforts to include more realistic applications, with an emphasis on the ways in which mathematics is used in daily and professional life" (p. 7). Second and broadening the first development from applying to learning mathematics, there is an increasing use of examples from reality as integral part of an instruction theory for mathematics. For instance, in Realistic Mathematics Education (RME) (Freudenthal, 1973; Gravemeijer, 1994, 1999, 2004; Van den Heuvel-Panhuizen, 2000; Van den Heuvel-Panhuizen & Drijvers, 2014) contexts, models and representations play an important role in the educational process. The central idea in RME is that students should be supported in reinventing mathematics with the support of the teacher and the curriculum materials (e.g. textbook). The starting points of such reinvention processes should be experientially real for the students. Problems situated in every-day life contexts often fulfil this requirement. Hence in RME, situations from real life are not just used to prepare students for solving applied problems. The main function of real-life situations in RME is to offer a conceptual basis for reinventing the mathematics the students are to learn. In relation to this, Freudenthal's (1983) didactical phenomenology suggests to look for phenomena that—as he puts it—"beg to be organized" by the mathematical procedures, concepts or tools one wants the students to (re)invent. In addition to this, Treffers (1987) recommends a broad phenomenological exploration in order to incorporate various inroads to the mathematical procedures, concepts or tools under consideration. Third, in mathematics education research there is an increasing focus on problem-solving and modelling (Blum, Galbraith, Henn, & Niss, 2007; Burkhardt, 2006; Kaiser, Blomhøj, & Sriraman, 2006; Lesh and Zawojewski, 2007; Schoenfeld, 1992; Sriraman, Kaiser, & Blomhøj, 2006). Schoenfeld (2014) signals a reframing of what it means to understand mathematics:

> At the core of that reframing is the notion of mathematics as a sense making activity—that learning mathematics entails developing deep understandings of certain culturally and historically transmitted ideas, and employing those ideas in ways that reflect the perception of objects and relations, their mathematization, and the meaningful use of mathematical symbols in the service of solving problems. (p. 498)

The aforementioned developments concerning applicability, instruction, and problem solving and modelling can be seen as branches of a larger tree that represents the delicate relationship between reality and mathematics (education) and subsequently the way that reality is represented in classroom practice. As a consequence of these developments, in classroom practice we see a great variety of examples and problems from real life, which are used as tasks. The predominant form used for these tasks is the

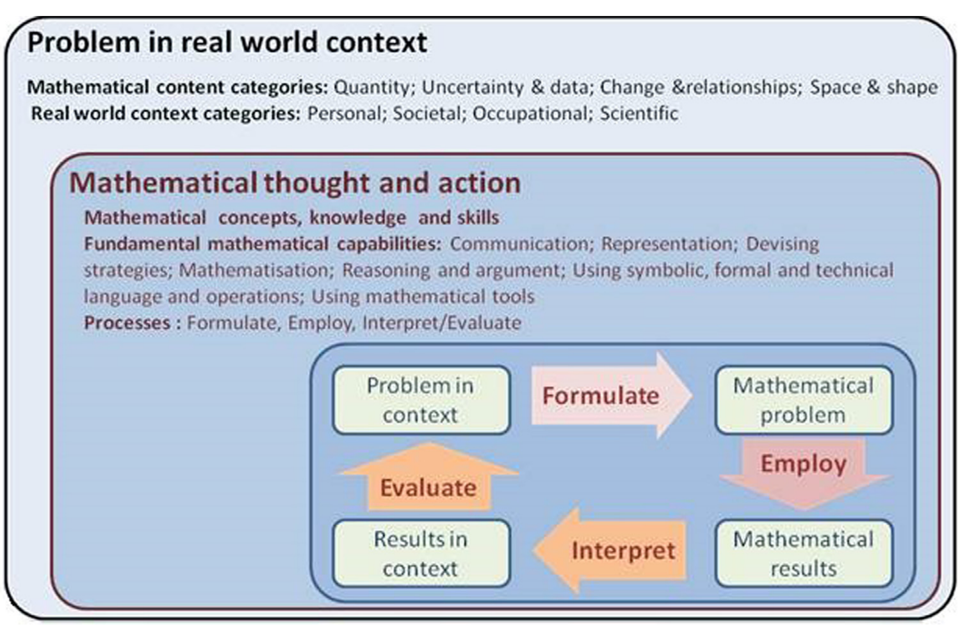

Fig. 1. A model of mathematical literacy in practice, according to OECD (PISA).

word problem, which uses above all a descriptive representation of reality (Madison and Steen, 2008; Sriraman et al., 2006). This use of word problems however is not without difficulties.

In the next section we will show the difficulties that are encountered, when using word problems in classroom situations; which ideas for counteracting those difficulties are conveyed in the literature; and a potential alternative to be designed.

2.2. Difficulties with sense-making in solving word problems

In the literature on problem solving and modelling several diagrams are used to visualize the process of solving word problems (Blum et al., 2007; Burkhardt, 2006; OECD, 2013; Verschaffel et al., 2000). In Fig. 1 the OECD (2013) schema is shown.

The aforementioned diagrams have in common that they show a variety of cognitive activities that are possible and necessary to solve quantitative problems from everyday life. It is assumed that somewhere in the process the solver must formulate a (mental) mathematical problem from the problem situation, which gives him the opportunity to employ the relevant mathematical reasoning and calculations towards a mathematical solution of the problem. And typical for this kind of problems, the solver has to interpret the outcome of the mathematical situation to make sense of it in the perspective of the original problem. We can speak of a problem solving mindset if a student fully engages in the activities formulated in these diagrams. However, each step also stands for possible barriers in the problem solving process.

Many studies on students' behaviour in solving word problems report that the steps of understanding the situation are often superficially executed by the students (Verschaffel et al., 2000). Other difficulties with word problems arise when students disregard possible constraints imposed by reality when they experience unfamiliarity with the situation at hand (Cooper and Harries, 2003; Lave, 1992; Reusser and Stebler, 1997; Verschaffel et al., 2009), or lack the proper meta-cognitive skills for solving the problem (Caldwell, 1995). The most reported difficulty is that students base their analysis and calculations on a loose association of certain salient quantitative elements of the problem situation with certain mathematical operations (Verschaffel et al., 2000).

Students seem to merely focus on the right hand side vertical step of the diagram in Fig. 1 and on the outcome of the calculation, and they seem to value these outcomes higher than the "realism" of the outcome. This student behaviour is an indication that students see a word problem as a "school-maths" problem disguising some arithmetic algorithms, and not as a representation of a real-life problem that has to be understood. This behaviour stems from a strong calculational orientation in mathematics classrooms (Thompson, Philipp, Thompson, & Boyd, 1994), whereby the focus is on procedures and operations and not on explaining and reasoning. In the calculational orientation, quantitative problems seem first and foremost to be used to train students in procedural fluency. From this perspective, the representation of the problem situation is not a particularly important aspect, and the word problem chosen is typically straightforward, scarcely hiding the aim of merely being a means of training for a certain type of calculation. The phenomenon that students often fail to note the meaninglessness of the stated problems and belief that every word problem has to be solved by a single numerical answer, was coined by Schoenfeld (1991) as "suspension of sense-making". This phenomenon is further investigated in the monograph "Making Sense of Word Problems" (Verschaffel et al., 2000).

The persistence of the perceived difficulties cannot exclusively be ascribed to student attitude. The literature contends that this student behaviour is likely to be reinforced by teachers' approaches that tend to emphasize the mathematical structure of the problem, rather than the contextual aspects (Depaepe, De Corte, & Verschaffel, 2010). Gravemeijer (1997) and Yackel and Cobb (1995) have pointed out that the typical students' and teachers' behaviour of focusing on outcomes is widespread, and persistent and based on established implicit or explicit socio-mathematical norms in classroom culture. Furthermore, they argue that one norm stands out: In school mathematics solving word problem is about finding and performing the right calculation.

The mentioned difficulties give rise to serious concerns whether students can show their full potential of solving quantitative real-life problems when confronted with word problems in an assessment or evaluation situation. Hence, it is proposed that an alternative for word problems should be considered. And that such alternative should be tested in practice and that effects of such alternative should be measured in a systematic way.

2.3. Counteracting difficulties with word problems

Over the past decades many attempts have been made to formulate quantitative problems, instruct students and create situations in such a way that the calculational approach and suspension of sense-making around word problems would be counteracted, or avoided. Verschaffel, De Corte, and Lasure (1994) whose work has been reproduced in many countries, researched the effect of using paired standard and 'problematic' items, whereby for the problematic items the students were almost forced to take into account the realistic constraints of the situation. The conclusion was that students have such a strong tendency to exclude realistic considerations that 'problematizing' the word problem was not enough to trigger realistic considerations to any great extent. Reusser and Stebler (1997) researched the effect of alerting students explicitly to the need to weigh realistic considerations, and of suggesting meta-cognitive activities, such as making a sketch or studying the picture. They found a small positive effect for these measures, but their main conclusion was that the classroom culture of not taking into account realistic considerations and hastily looking for the right operation and the right answer persists. Wyndhamn and Säljö (1997) studied the effect of increasing the authenticity of the experimental setting by actually bringing in concrete materials that played a role in the given problems. They found that students scored significantly better in the more authentic settings, which they saw as evidence that the presentation of the problem was the main cause for the students' unrealistic answers. Similar research was done by DeFranco and Curcio (1997) who compared the results on a written problem – 328 Senior citizens are going on a trip. A bus can seat 40 people. How many buses are needed so that all the senior citizens can go on a trip? – with a real-life simulation of this problem where the students actually had to call the bus operator to order buses. They found a reduction in the reluctance of students to take realistic considerations into account. Cooper and Harries (2002, 2003) singled out a problem from an English national test about the number of times an elevator had to go up to transport a certain number of people. They made an alternative problem where students were encouraged to seriously consider the practical constraints and the assumptions in the calculations. They found that around a quarter of the children proved able and willing to offer reasons for their choices made. Dewolf, Van Dooren, and Verschaffel (2011) examined the effect of presenting the same fair sharing problem in different settings – mathematics class versus religion class – and found that the context of problem solving has an important influence on the interpretation and solution of the problem: a substantial number of students in the religion class combined and weighted a larger variety of criteria for the problem statement, whilst students in the mathematics class limited themselves mainly to one criterion for fair sharing.

In research on alternatives to word problems, other researchers and practitioners have asked whether the problem of suspension of sense-making can be counteracted by changing the problems, or adding incentives to students to make realistic considerations (Bonotto, 2007, 2009; Frankenstein, 2009; Lave, 1992; Zevenbergen & Zevenbergen, 2009). They advocate the creation of actual real-life situations in mathematics lessons, which could feel more authentic to students, to keep students more in a problem solving mindset. Bonotto (2007) argues for encouraging students to analyse mathematical facts, embedded in appropriate 'cultural artefacts', such as supermarket bills, 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 not a widespread dissemination of such practices. In our alternative we decided to change the representation of the problems situation from descriptive to more depictive to get it closer to the real-life situation.

2.4. Perspectives on the use of depictive elements

Recent studies by Dewolf c.s. (Dewolf, Van Dooren, Ev Cimen, & Verschaffel, 2014; Dewolf, Van Dooren, Hermens, & Verschaffel, 2015) and earlier studies of Elia, Gagatsis, and Demetriou (2007) and Elia and Philippou (2004) addressed the effect of adding decorative, representational, and informational pictures to mathematical word problems, and they found small or no significant effects. These studies encouraged us to look into the designing process for images closely connected to the real-life situation that was to be represented. So, in our approach, the problem situation was represented with images from real life, in the form of photographs, headlines from newspapers, and handwritten notes (see Fig. 2). We hypothesised that the use of images from real life would increase the association with real-life situations and augment the chance that the solver is "solving a truly engaging dilemma" (Verschaffel et al., 2000, p. 44) and therefore decrease the suspension of sense-making and the strong calculational focus.

Recent studies also investigated the relation between the use of pictorial elements by students during the problem solving process and their performance on word problems (Boonen, 2015; Hegarty, 2004; Krawec, 2014; Van Garderen, 2006; Van Garderen and Montague, 2003). These studies show that students perform better when they make use of relevant visual or schematic elements in their problem solving process. However, this only held when the pictorial elements of the students were relevant for and consistent with the mathematical model needed to solve the problem. When students used less relevant pictorial drawings the effect on their performance was negative.

Furthermore, a supporting argument for the change in the direction of more depictive representations can be found in research on the linguistic issues students encounter in solving word problems. For example, research studies investigating the role that language plays in the actual solving of word problem solving (Sepeng and Webb, 2012), for instance through discussions, show that linguistic development of students have an effect on their problem solving skills and on the interplay between contexts and sense-making (Sepeng, 2013). This relation between linguistic development and mathematical problem solving performances is corroborated by the 2012 PISA studies that show that there is a latent correlation of 0.85 between reading and mathematics performances (OECD, 2014).

These studies provided us with arguments that designing a specific, relevant and close to real life form of visual representation of the problem situation might help students in making sense of the problem, adopt a relevant mathematical model, and interpreting the outcomes of the mathematical operations. This led to a set of choices we made in designing the tasks for the instrument.

2.5. Our choices for an alternative approach

In our research we sought for an alternative for word problems that could be used in assessment and evaluation instruments, and at the same time portray real-life situations, that feel authentic to students. Authenticity in education is a widely debated subject, that dates back to the writings of Dewey (1916). Palm (2009) argues in his Theory of Authentic Situations "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 situations that are simulated" (p. 9). From Palms framework (Palm, 2009) the aspects realism and presentation of the problem were used in our alternatives to make the represented problem situation feel more authentic to students. The other aspects like event and circumstances were not changed in our particular comparison. We contend that those first two aspects contributed to a higher representativeness of the tasks which arguably could have a positive effect on the problem solving quality of the students.

From a cognitive psychological perspective Schnotz et al. (2010) distinguished between descriptive and depictive representations, to each of which they ascribed specific representational and inferential powers. In a much similar approach as was used in the problem solving schematic of Fig. 1, Schnotz (2002) stated that for solving a quantitative problem, a task-oriented construction of a mental mathematical model has been necessary. Their line of reasoning was that depictive representations could better support students to make a relevant mental mathematical model of the situation. They further stipulated that depictive representations

3A

On the highway you see a road sign
which says it is still 39 km to Amsterdam
On this highway the maximum speed is
100 kilometer per hour.

How many minutes will it take to reach Amsterdam with this speed?
☐ **minutes**

3B

How many minutes will it take to reach Amsterdam with this speed?
☐ **minutes**

Fig. 2. Example of word problem (left) and image rich numeracy problem (right).

had a high inferential power and were in many cases close to the form of mental mathematical models: sketches; diagrams; and/or drawings in that they captured the essence of the problem (Schnotz et al., 2010).

Based on this reasoning we inferred that using images from real-life situations could support students in their problem solving task.

2.6. Design of the instrument and validity issues

Informed by the aforementioned research on word problems, cognitive psychological perspectives on problem solving, and linguistic considerations, we came to the design of an instrument to measure the effect of changing the representation of the problem situation on students' performance. Aiming for large scale use, we designed the instrument in such a way that it could simultaneously serve as a diagnostic test for a, in the Dutch situation upcoming, nationwide examination on numeracy. In the design of the instrument we carefully considered issues of validity, according to the AERA/APA/NCME recommendations (Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association the American Psychological Association and the National Council on Measurement in Education, 2014).

For measuring the effect of changing the representation of the problem situation on students' performance, we designed the instrument as a randomized controlled trial, allowing to analyse the data with more sophisticated limited dependent variable models, in order to establish the plausibility of the effect measured and to interpret the effect of possible interdependent variables. For the diagnostic use of the instrument we focused on the validity of the content of the test. In the method section we detail the evidence we gathered to support the validity of the intended instrument uses and interpretations.

2.6.1. Context of the use of the instrument

For international readers we provide information on the context, in which the instrument was 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 by law (Ministerie van OCW, 2009). In this framework six levels are formulated (see Table 1), and four (out of the six levels) are assessed in nationwide examinations at the end of primary and secondary education.

The F-levels focus mainly on so-called functional mathematics and are assessed with tests consisting of 45–55 mostly word problems (Cito, 2015). In the framework, four content domains are formulated: numbers; proportions; geometry & measurement; and relations (tables, diagrams, graphs, formulas, etc.). A diagnostic tool for these upcoming nationwide examinations would be supportive for teachers and students, and insights into the effect of changing the representation of the problem situation is likely to contribute to the body of knowledge on how to assess the Dutch LaNF-framework in a research-informed way (Hoogland and Stelwagen, 2011).

2.6.2. General layout of the final instrument

The final instrument is a web-based numeracy test of 24 items, very similar in content and layout to the nationwide examination (Cito, 2015). The 24 items are diagnostic for the level 2F of the Dutch LaNF. Of these items, 21 items are randomly presented in one of two versions: word problem; or image-rich numeracy problem, which means that the problems are equivalent regarding the content and the level of mathematical knowledge and skills needed to solve the problem. The versions only differ in the way the problem situation is presented to the participant (see Fig. 2). This last feature of the instrument makes it possible to measure the effect of changing the representation of the problem situation on students' performance in a randomized controlled way. In the method section we described the activities undertaken to counter threats to validity of both proposed uses.

3. Method

The activities undertaken in the designing of the instrument and the validation of its different uses are presented in Table 2. For the use as a measuring tool on the difference in students' performance related to the representations of the problem situation, we foremost focused on the equivalence of the word problems and the image-rich problem. For diagnostic use we focused on the construct and content validity of the items. In order to counter threats to reliability an important aspect was computer scoring the students' solutions, to guarantee that for each participant the scoring was consistent.

Next to activities to provide evidence for construct and content validity we argued for criterion validity with measures obtained in the test run. However, we used these measures with caution, because we are aware that the use of criterion validity is also a topic of discussion among psychologists (Borsboom, Mellenbergh, & van Heerden, 2004). In the next paragraphs we explain the activities undertaken in more detail.

3.1. The process of designing

The process of designing started with the selection of 40 relevant word problems that were used in recent years in Dutch textbooks and tests, which were developed to teach or asses the numeracy 2F level in the LaNF. Additional selection criteria were that the problems were dealing with a real, perceived as real, or at least imaginable problem from daily life. The selected problems were specifically aiming at level 2F of this framework, and more particularly evenly spread over the domains of *numbers*, *proportions* and *geometry & measurement*. Items on the domain *relations* were later added to the instrument to allow for the possible use as a diagnostic test. We limited ourselves to one version for those items, because in the domain relations hardly any word problems without a visual element could be found. The selected problems for

Table 1
Overview of Numeracy Levels and Moments of Examination of Dutch Literacy and Numeracy Framework.

Level	Description	Moment of Nationwide Examination
1F	basic level	end of primary education
1S	advanced level	end of primary education
2F	basic level	end of secondary general and vocational education, lower tracks
2S	advanced level	not tested
3F	basic level	end of secondary general and vocational education, higher tracks
3S	advanced level	not tested

Note. F stands for Fundamenteel Niveau [Basic Level], S stands for Streefniveau [Advanced Level].

Table 2
Overview of activities undertaken in design of the instrument to counter threats to validity and reliability.

Phase of development	Number of research activity	Description of Research Activity	To counter threats to:
Design	1	Selecting 40 existing items around level 2F of the LaNF.	content validity
	2	Designing 40 alternatives and gathering comments on quality of 40 paired problems by 13 experts.	content validity
	3	Estimation of levels 2F of 40 revised paired problems by eight experts.	construct validity
Validating the diagnostic use	4	Creating a web based version similar to the nationwide examination in content and layout.	construct validity
	5	Relating the items to the LaNF and spreading the items evenly over the domains of the LaNF.	content validity
	6	Performing a test run with over 7000 participants.	feasibility
	7	Checking for internal consistency of the items with measures of the classical item response theory.	criterion validity
	8	Checking correlation of scores on both versions	content validity
Validating the measurement of changing the representation of the problem situation.	9	Checking 40 revised paired problems on equivalence of paired items by eight experts.	content validity
	10	Programming random representations of the problem situation in 21 items and presenting them in random order in the instrument.	construct validity and reliability
	11	Computer scoring students' solutions.	reliability
	12	After test run: checking for correlation between scores on both versions.	content validity
Composing the final instrument	13	Combine results from all above to construct the final instrument	

the other domains had as common characterization: the problem situation was described in words, without the use of illustrations or photos. The selected problems where all standardized in the same format: problem situation; problem question; and answering box (see Fig. 2).

For all 40 selected word problems an alternative version was designed, in which the descriptive representation of reality in the problem situation was replaced, as much as possible, by a depictive representation of the proposed reality. As depiction we choose one or more images from reality, mostly photographs, with little redundancy, whereby as little language as possible was used. These image-rich problems where standardized in the same format (see Fig. 2) as the word problems.

The problem question and the answering box were identical. Only the representation of the problem situation was changed. The word problems with a descriptive representation of the problem situation we called the A-version; and the image-rich problems with a depictive representation of the problem situation we called the B-version, resulting in 40 paired problems. In Table 2 these research activities are numbered 1, 4 and 5.

3.2. Validation activities by experts

Experts in mathematics education played an important role in the process of content validation. The aim was to design paired problems that only differed in the representation of the problem situation, leaving all the other possible variables the same. For valid conclusions on the effects of changing the representation of the problem situation, the pairs of problems had to be, apart from representation of the problem situation, otherwise equivalent. In a first cycle to improve the items, the 40 paired problems were openly discussed by 13 experts on quality, relevance for the level 2F and equivalence between A-version and B-version. This lead to an improved set of 40 paired problems.

In a second cycle the improved set was presented to eight other Dutch experts in the field of mathematical literacy and numeracy, with the following questions:

(i) "Do the two versions of the problem test the same mathematical knowledge and skills?"
(ii) "If the two versions test the same mathematical knowledge and skills, are they testing on the same mathematical level?"
(iii) Give for each problem the estimated level on a five-point scale: too easy for level 2F, easier than level 2F, level 2F, more difficult than level 2F, too difficult for level 2F.

The experts were explicitly asked to disregard their estimation of the effect of changing the representation of the problem situation on the difficulty of the problem. This expert group consisted of teachers, teacher educators, researchers and test constructors in the domain of mathematics and numeracy education. These experts sent in their results anonymously. We as designers decided after discussion that a pair of problems was acceptable for the instrument, if the first question was answered with "Yes" by a minimum of all but one of the experts. The answers to the second and the third question made it possible that in the final instrument the selected problems were spread evenly around level 2F of the LaNF. In Table 2 these research activities are numbered 2, 3, 5 and 9.

3.3. The test run of the instrument

The test run of the instrument took place in parallel (with respect to time) with the activities by the experts. From the list of

Table 3
Number of Participants in Test Run ($n = 7434$).

	Primary education (grade 5–6)	Secondary education (grade 1–6)			Total
		VMBO (pre-vocational)	HAVO/VWO (general)	MBO (vocational)	
N	172	3796	2838	348	7434

Note. VMBO is the pre-vocational track in secondary education, HAVO/VWO is the general track in secondary education, MBO is the vocational track in secondary education. Participants are representative to proposed future use of the instrument.

40 paired problems 21 paired problems were selected considering a good spread on the domains numbers, proportions, geometry & measurement and the aforementioned three items were added on the domain of relations. In this way the test would qualify as a complete 2F test according to the LaNF. The 24 items (21 items in two versions, and three items in one version) were programmed as a web based digital test. From experience with earlier pilot tests of the LaNF we estimated that the 24 problems could be solved in about 50 min. The selection of the problems for the test run took place in time before the validation by experts, so the selection was done by the designer with the following criteria: evenly spread over the domains numbers, proportions, geometry & measurement, and evenly spread over a bandwidth of levels around 2F.

3.3.1. The participants for the test run

The participants for the test run were selected by inviting schools to participate. We looked for participants from a broad range of tracks and levels to gather information on the feasibility of the items and the test as a whole. The distribution of the participants in the test run of the instrument around the levels of education is shown in Table 3.

In the test run of the instrument 7434 students from 63 different schools participated in an age range of 11–18 years old. The schools in the test run of the instrument were geographically spread around the Netherlands and delivered students from all levels of education, which gave a good representation of the proposed use of the instrument in the future.

3.3.2. Conducting the test

For each participant a test was generated by randomly choosing 12 items to be presented with a descriptive representation of the problem situation (A-version) and the other 12 to be presented with a depictive representation of the problem (the B-version). The order of the problems presented to the participants was again randomized. To explain, for instance a test could look like the following: 4B, 5B, 9B, 1A, 3B, 6A, 12B, 23A, 24B, 19A, 15A, 17B, and so on. Each problem in the web-based test was presented as a screen filling problem. The question was posed at the bottom of the screen. Below the question the numerical solution to the problem could be entered. For solving the problems an on-line calculator was allowed. For the total test a time limit of 60 min was set, to make it possible for the students to also answer a few short additional questions. All solutions to the problems were numerical values. Participants typed the numerical solutions into an empty answering field. The answers were evaluated and scored by the computer.

Table 4
Problems, expert judgments, average expectations and selection for the test run and the final instrument.

Problem	Same knowledge		Same level		Estimated level		test run	final instrument
	N	Yes	*N*	Yes	M(A)	M(B)		
AEX-index	8	100%	8	100%	3.88	3.88	P17	v
Endive	8	100%	8	100%	3	3		v
Coughing syrup	8	100%	8	100%	3.25	3.25		v
Coffee cups	8	100%	8	100%	3	3	P2	
Scale model	8	100%	8	100%	3.13	3.13	P20	v
Winter tiers	8	100%	8	100%	3.75	3.75		v
Chicory	8	100%	8	100%	2.25	2.25		
Bathroom	8	100%	8	88%	3.63	3.63		v
Baking tin	8	100%	8	88%	3	3		v
Budget cuts	8	100%	8	88%	3.63	3.63	P18	v
Public debt	8	100%	8	88%	3.88	4		v
Gas usage	8	100%	8	75%	2.75	2.75	P5	v
Driving time	8	100%	8	75%	3.25	3.38	P7	v
Apples	6	100%	7	71%	2.86	2.71	P1	v
Petrol	8	100%	8	63%	3.75	4		
Offer on buying	8	100%	8	50%	3.13	2.88	P10	v
Groceries	8	88%	8	88%	2.5	2.5	P13	v
Lawn fertilizing	8	88%	8	88%	3.13	3.13		
Hamburger	8	88%	8	88%	3.13	3.13		v
Icelandsic Crowns	8	88%	8	88%	3.63	3.63	P16	v
Bedroom tiles	8	88%	8	88%	3.38	3.38		v
Water bottles	8	88%	8	88%	2.88	2.88		v
Double glazing	8	88%	7	86%	3	3.29		v
Bank balance	8	88%	8	75%	2.13	2.25		
Weed control	8	88%	8	75%	3	3		
World cities	8	88%	8	75%	3.71	3.57		
Swimming pool	8	88%	8	75%	3.13	3.13		
Bonbons	8	88%	8	63%	3.13	3.25		v
Music Songs	8	88%	8	63%	3.63	3.63	P14	
Water usage	8	88%	8	50%	3.75	3.38	P4	
Recipe	7	86%	7	86%	3.13	3.13	P15	v
Carpet	8	75%	7	71%	2.71	2.75	P6	
Purchase laptop	8	75%	8	63%	3.13	3	P8	
Cubes	8	75%	8	50%	4	4	P21	
Art work	8	75%	8	50%	3.14	3.13	P12	
Elephant	8	75%	8	38%	4.25	4.25	P3	
Book shelves	8	63%	8	63%	3.25	3.13	P9	
Packing box	8	63%	8	50%	3.71	3.43		
Thermometer	8	63%	7	43%	2.38	2.38	P19	
Photo	8	63%	8	25%	4	4	P11	

Note. *N* is number of experts, Yes is percentage of affirmative answers of the experts to question1 respectively question 2. M(A) is mean expert estimation of the word problem levels on a 5-point scale referenced to 2 F level. M(B) is the mean expert estimation the image-rich problem levels on a 5-point scale referenced to 2 F level. The column *test run* shows the problems used in the test run. The column *final instrument* shows the problems selected for the final instrument.

In Table 2 these research activities are numbered 6, 10 and 11. In the results section the outcomes of the remaining research activities in Table 2 are discussed, to show how we countered threats to the validity of the two proposed uses of the instrument.

4. Results

4.1. Results of the design process

The resulting items of the first cycle of the design process can be found under open access (Hoogland & De Koning, 2013): *Overview of 40 paired problems, with three questions for eight experts*. The results on the three questions posed in the second cycle can be found in Table 4.

From the criterion that a pair of problems was acceptable for the instrument if the first question ("Do the two versions of the problem test the same mathematical knowledge and skills?") was answered with "Yes" by a minimum of all but one of the experts, 30 of the 40 paired items were confirmed as eligible for the final instrument (See in Table 4 the items above the bold line). To spread the items evenly over the domains another nine items were discarded.

For the 21 problems that were selected for the final instrument we found as mean estimation (with standard deviation in parentheses) of the level 3.23 (0.37) for the word problems, and 3.25 (0.39) for the image-rich-problems. For an average estimation on a 5-point scale this is a good indication for equivalence. In Table 2 these research activities are numbered 5 and 8.

4.2. Results from the test run of the instrument

After the test run of the instrument the test as a whole was evaluated by analysing the average good scores of the items, the item-rest-correlations (RiR), and the correlation between the numbers of words, visual elements and test scores (see Table 5).

The average good scores of the items range from 0.10 to 0.90, with the exception of P4, which was not included in the final instrument. The item-rest correlations are all positive and higher than 0.25. For the test of the 21 items we found Cronbach's $\alpha = 0.82$ and for the test of the 24 items we found Cronbach's $\alpha = 0.83$, which indicates a good internal consistency (Kline, 1999). For the diagnostic use of the instrument we considered that this was indicating sufficient criterion validity. We also checked the correlation of the scores on A-versions and B-version and found $r = 0.98$, which contributes to both the content validity in measuring the same construct as well as to the criterion validity in adding to the internal consistency of the instrument. In Table 2 these research activities are numbered 7 and 12.

Furthermore in the test run, we investigated the relation between the number of words used in the representation of the problem situation in the A-version and the scores on the A-version of the items. We found a moderate correlation $r = -0.45$ indicating that the word problems with more words were more difficult. Likewise we found $r = -0.34$ for the correlation between the number of visual elements in the representation of the problem situation of the B-version and the scores on the B-version, indicating that the problems with more visual elements were more difficult. In this test run we found no correlation ($r = 0.00$) between the reduction of the number of words and the difference in performance between the two versions. In results obtained with the final instrument with validated items, this will be investigated in more depth.

4.3. The resulting final instrument

In constructing the final instrument another nine of the remaining 30 paired problems were discarded based on the following criteria: evenly spread over the domains numbers, proportions, geometry & measurement, evenly spread over a bandwidth of levels around 2F and discarding some items for which the underlying mathematical structure were almost identical, for example "chicory" and "endive", and "coffee cups" and "water bottles".

And finally, with the combined results from the expert validation and the test run of the instrument the final instrument was composed and constructed as a web-based test. Ultimately 21 paired items on the domains numbers, proportions and geometry & measurement (see Table 4), and three additional items on the domain relations were selected for the final instrument. The three

Table 5
Scores on both versions, Item-Rest-Correlations (RiR), word count (A-version) and number of visual elements (B-version) from test run ($n = 7434$).

item	item score (A-version)	item score (B-version)	RiR (A- + B- version)	word count (A-version)	number of visual elements (B-version)
P1	0.69	0.75	0.42	17	2
P2	0.55	0.56	0.42	15	2
P3	0.28	0.14	0.26	33	3
P4	0.09	0.13	0.28	37	3
P5	0.52	0.49	0.34	21	3
P6	0.77	0.82	0.36	42	2
P7	0.29	0.26	0.50	29	2
P8	0.83	0.84	0.32	17	2
P9	0.60	0.49	0.42	32	2
P10	0.78	0.83	0.43	28	2
P11	0.13	0.22	0.41	43	2
P12	0.42	0.48	0.44	22	2
P13	0.90	0.89	0.25	11	2
P14	0.26	0.27	0.41	17	2
P15	0.84	0.86	0.38	37	1
P16	0.43	0.39	0.51	26	2
P17	0.28	0.28	0.41	28	2
P18	0.16	0.18	0.33	31	1
P19	0.85	0.76	0.41	20	1
P20	0.47	0.51	0.48	23	2
P21	0.14	0.10	0.32	70	2

Note. Items P1, . . . , P21 can be found in Table 4. Item score is average good score. RiR is item-rest correlation Words and number of visual elements are counted in the representation of the problem situation. Cronbach's $\alpha = 0.82$ for these 21 items.

additional items did not play a role in the eventual analysis of the results. In Table 2 this research activity is numbered 13.

The instrument was made available as an on-screen test an any computer connected to the internet. When a student participates in the test a personal activation code is provided to start up the digital test of 24 problems and afterwards an optional additional short digital questionnaire is delivered to each participant to collect the following additional data: school level, grade level, gender, ethnicity, age, and last received math grade (test or student report). All data are recorded anonymously in a research database. The design of the instrument makes it possible to compare the performances on word problems and the performances on image-rich problems in a controlled randomized way. In October and November 2011 the instrument was used in the Dutch context with a sample of over 32.000 students (Hoogland, et al., submitted). With this article the English version of the instrument is made available under open access (Hoogland & De Koning, 2013).

5. Discussion

An important goal of mathematics education is fostering students' ability to use mathematical knowledge and skills to solve problems from daily life. Worldwide the practical use of mathematics is seen as one of the justifications for mathematics education (Kilpatrick, 1996). To reach this goal there is a need for adequate teaching materials and assessment tools. In those materials and tools quantitative problems from daily life are typically represented. For many decades, even centuries, those representations have been dominated by descriptive representations of reality: the resulting problems have been numeracy word problems that students had to read and make sense of. And although in educational research the use of word problems has been problematized (Gellert and Jablonka, 2009; Roth, 2009; Verschaffel et al., 2009), in classroom practice word problems are, without much discussion, seen as accepted parts of mathematics education, although the use of (too much) language is seen by many teachers as an additional difficulty, especially for low performing students. According to our findings, the dilemma of using word problems versus using real-life situations for teaching and assessing students in solving problems from daily life seems to have been "overcome"; a third way seems possible. The representation of a problem situation can get much closer to the real-life problem situation by using photographs. This is a potential way to bring in more perceived authenticity into the mathematics classroom, without automatically introducing the practical constraints that occur by introducing real artefacts and simulations into the classroom. The aim of our research was to show a feasible alternative, by designing more image-rich alternatives for existing word problems. But before we could conclude that such an alternative for word problems was feasible in the mathematics classroom, we had to gain more insights into the effect of such an alternative (in the way the problem situation is represented) on students' performance. To measure this effect, we needed an instrument that could validly measure such an effect. In this study we described the design of such an instrument that can measure the effect of changing the representation of the problem situation on the performance of the students in a controlled randomized way. This made it possible to analyse the effect also in relation to other possible intervening variables.

To validate the instrument uses we took three pathways: firstly, we designed the alternative items and referenced them with the Dutch Literacy and Numeracy Framework; secondly, we had two panels of experts to comment the problems used in the instrument, and to validate the equivalence of the selected word problems and the alternative image-rich numeracy problems. Thirdly, we took a test run with the instrument to see, if the condition of a controlled randomized trial could be met. These activities around validations were carried out successfully, so that the final instrument can be used to validly measure the effect of changing the representation of the problem situation.

Our instrument also has limitations. The number of problems used for the instrument was limited. Further study is necessary to get a better view of which problems in particular were sensitive to the change of representation of the problem situation, which were not, and why.

In this article we contended that, despite the limitations, the instrument can reliably measure the difference in performance on two representations of the problem situation. There is, however, still a long way to fully understand which are the underlying factors that actually could cause a difference in performance. In the analysis of the results the effect of task characteristics, like the content domain the task belong to, might be worth investigated. In the analysis we intend to take task characteristic variables such as wordiness of the word problem or the number of used images for the image rich representation as interdependent factors to identify the underlying patterns. This could lead to replications of the study with more and more specifically designed items.

At the same time more qualitative research should be done that focus on the actual behaviour and thought patterns of students when solving the problems, for instance by using thinking aloud protocols or stimulated recall. This could shed more light on the intricate relation between use of language, problems solving capabilities, and sense-making.

6. Conclusions

The instrument designed and validated in this study is likely to contribute to further knowledge and insights in terms of how the representation of reality in contextual mathematical problems can affect the performance of students, and as a consequence the outcomes of their assessment and the subsequent conclusions.

In follow up studies this instrument was scheduled to be used in a large scale test of students (over 30,000 participants), in which more data on the effects of changing the representation of the problem situation were gathered, also in relation to other measured characteristics of the students and of the items. By using a latent variable model and analysis, conclusions about the expected differences can be made with a good degree of validity. Apart from establishing the effect of changing the representation of the problem situation on students' performance, we hope to find better and more concrete indications why the differences in effect appear, and why effects appear with certain type/genre of problems and not with other types/genres, for instance related to the wordiness of the word problem or the number of pictures used in the image rich tasks.

In our endeavour to get a better understanding of the effects of the representation of real-life problems in mathematical contextual problems, we have designed, trialled and tested an instrument that can give an indication of the effect of changing particular representations (from word to picture), and can also provide the opportunity to analyse this effect systematically in relation to other variables that influence students' performances in solving problems from daily life. We invite researchers to use the English or other language translations of the instrument to further investigate the effects of representation of reality in mathematical contextual problems.

Acknowledgement

The research in this paper 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

Modelling and applications in mathematics education – the 14th ICMI study. In W. Blum, P. L. Galbraith, H.-W. Henn, & M. Niss (Eds.). New York, NY: Springer.

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.

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*. Vrije Universiteit Ph.D.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *4*(111), 1061–1071.

Burkhardt, H. (2006). Modelling in mathematics classrooms: reflections on past developments and the future. *ZDM*, *38*(2), 178–195. http://dx.doi.org/10.1007/BF02655888.

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

Cito (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.

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.

Daro, P. (2013). *Phil Daro - against answer getting*. [Video] Retrieved from https://vimeo.com/79916037.

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.

Dewey, J. (1916). *Democracy and education: an introduction to the philosophy of education*. New York, NY: Simon & Schuster.

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.

Elia, I., & Philippou, G. (2004). The functions of pictures in problem solving. In M. Johnsen Hoines, & A. Berit Fuglestad (Eds.), *Proceedings of the 28th conference of the international group for the psychology of mathematics education:* (vol. 2. pp. 327–334).Bergen, Norway: Bergen University College.

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.

Freudenthal, H. (1973). *Mathematics as an educational task*. Dordrecht, the Netherlands: Reidel.

Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures*. Dordrecht, the Netherlands: Reidel.

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.

Gravemeijer, K. (1994). *Developing realistic mathematics education*. Utrecht: CD-ß Press/Freudenthal Institute.

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.

Gravemeijer, K. (1999). How emergent models may foster the constitution of formal mathematics. *Mathematical Thinking and Learning*, *1*(2), 155–177. http://dx.doi.org/10.1207/s15327833mtl0102_4.

Gravemeijer, K. (2004). Local instruction theories as means of support for teachers in reform mathematics education. *Mathematical Thinking and Learning*, *6*(2), 105–128. http://dx.doi.org/10.1207/s15327833mtl0602_3.

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.

Hoogland, K., & Stelwagen, R. (2011). A new Dutch numeracy framework. In T. Maguire, J. J. Keogh, & J. O'Donoghue (Eds.), *Mathematical Eyes: a Bridge Between Adults, the World and Mathematics. Proceedings of the 18th International Conference of Adults Learning Mathematics—a Research Forum (ALM)* (pp. 193–202).Tallaght, Ireland: ITT.

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

Hoogland, K., De Koning J., Bakker, A., Pepin, B., Gravemeijer K., 2016. Descriptive versus depictive representation of reality in contextual mathematical problems: the effect on students' performances. (submitted).

Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association the American Psychological Association and the National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association: Washington DC.

Kaiser, G., Blomhøj, M., & Sriraman, B. (2006). Towards a didactical theory for mathematical modelling. *ZDM*, *38*(2), 82–85. http://dx.doi.org/10.1007/BF02655882.

Kilpatrick, J. (1996). Introduction to section 1: curriculum, goals, contents, resources. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education*Dordrecht, the Netherlands: Kluwer.

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.

De Lange, J. (1999). *PISA mathematics framework*. Paris, France: Organisation for Economic Co-operation and Development (OECD).

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.

Lesh, R., & Zawojewski, J. (2007). Problem solving and modeling. In F. K. LesterJr. (Ed.), *Second handbook of research on mathematics teaching and learning*Charlotte,NC: Information Age Publishing.

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.

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

OECD (2013). *PISA 2012 assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy*. Paris, France: OECD Publishing.

OECD (2014). *PISA 2012 results: creative problem solving: students' skills in tackling real-life problems, vol. V*, Paris, France: OECD Publishing.

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.

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.

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. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). Cambridge, UK: Cambridge University Press.

Schnotz, W., Baadte, C., Müller, A., & Rasch, R. (2010). Creative thinking and problem solving with depictive and descriptive representations. In L. Verschaffel, E. De Corte, T. De 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.

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

Sepeng, P. (2013). Issues of language and mathematics: contexts and sense-making in word problem-solving. *Mediterranean Journal of Social Sciences*, *4*(13), 51–61. http://dx.doi.org/10.5901/mjss.2013. v4n13p51.

Sepeng, P., & Webb, P. (2012). Exploring mathematical discussion in word problem-solving. *Pythagoras*, *33*(1), 1–8. http://dx.doi.org/10.4102/pythagoras.v33i1.60.

Sriraman, B., Kaiser, G., & Blomhøj, M. (2006). A brief survey of the state of mathematical modeling aroung the world. *ZDM*, *38*(3), 212–213. http://dx.doi.org/10.1007/BF02652805.

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.

Toner, P. (2011). *Workforce skills and innovation (OECD education working papers)*. . . (Retrieved from Paris, France) https://search.oecd.org/innovation/inno/46970941.pdf.

Treffers, A. (1987). *Three dimensions. A model of goal and theory description in mathematics instruction – the Wiskobas Project*. Dordrecht, the Netherlands: Kluwer.

Van den Heuvel-Panhuizen, M. (2000). *Mathematics education in the Netherlands: a guided tour*. Freudenthal Institute Cd-rom for ICME9.

Van den Heuvel-Panhuizen, M., & Drijvers, P. (2014). Realistic mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 521–525).The Netherlands: Springer.

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., De Corte, E., & 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., 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.

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

Words and worlds – modelling verbal descriptions of situations. In L. Verschaffel, B. Greer, W. Van Dooren, & S. Mukhopadhyay (Eds.). 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 (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.

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.