Hoogland K., & Pepin, B. (2016) The intricacies of assessing numeracy: Investigating alternatives to word problems. *Adults Learning Mathematics: An International Journal, 11*(2), 14-26

# The Intricacies of Assessing Numeracy: Investigating Alternatives to Word Problems

**Kees Hoogland**

SLO - Netherlands Institute for Curriculum Development

<k.hoogland@slo.nl>

**Birgit Pepin**

Eindhoven School of Education, Eindhoven University, the Netherlands

<b.e.u.pepin@tue.nl>

## Abstract

Word problems are often used to assess numeracy, despite the growing number of reports on difficulties students encounter with this genre of mathematical problems. These reports contend that a large number of difficulties are influenced by the way the problems are presented, that is, with verbal representations of the problem situations. These difficulties are said to be associated with a form of suspension of sense-making. In this study, conducted in the Netherlands, we investigated the effect on adult participants' performances of changing the representation of the problem situation, from verbal to image-rich. A controlled randomised trial was the main part of this investigation. Furthermore, we compared the results of adult participants with the results of a similar trial which was held with students from primary and secondary education. The study showed that adult participants' performances improved slightly with the change in representation, particularly on tasks in the content domain of measurement & geometry. These results were comparable with the results found of students from primary and secondary education, indicating that the effect is not related to age. The results could be of interest, however, for all practitioners involved in the work of numeracy task design.

Key words: numeracy, assessment, word problems

## Introduction

In most recent approaches in adult numeracy research, adult numeracy is defined as a complex, multifaceted, and sophisticated construct, incorporating the mathematics, communications, cultural, social, emotional and personal aspects of each individual in context (American Institutes for Research, 2006; Coben, 2003; Geiger, Goos, & Forgasz, 2015). As a consequence, learning and assessing numeracy in authentic situations is often advocated (Frankenstein, 2009).

A closer look at lesson or test materials used in numeracy education in many countries, however, reveals that most assessment materials consist of word problems or of items assessing procedural arithmetic skills. The same is the case in the Netherlands where, despite the country's high rankings in international comparisons, there are persistent complaints about the literacy and numeracy levels of young adults in vocational education and in the workplace. As a result, in 2010 a Literacy and Numeracy Framework (LaNF) was introduced in the Netherlands, with a compulsory numeracy examination at the end of the vocational educational tracks

(Hoogland & Stelwagen, 2012). After a lively debate on the assumed value of procedural skills for (young) adult learners, a compulsory numeracy examination has been implemented which consists of 45 mathematical problems of which 15 are strictly procedural problems and 30 arecontextual problems. Many teachers and mathematics educators have questioned the relevance of assessing vocational students this way, and they perceive a gap between the numeracy used by their young adult students in everyday life and (future) work, and numeracy as operationalised in the final examinations (Hoogland, 2006; Hoogland & Pepin, in press).

A study in 2011 and 2012 in the Netherlands focused on the idea that using image-rich numeracy problems contributes to bridging the gap between common classroom practice in numeracy and more sophisticated numeracy concepts (Hoogland, 2016). Part of this study was a controlled randomised trial with almost 32,000 primary, secondary, and vocational students, to investigate the effect on students' performance of changing the representation of the problem situation from verbal (word problem) to image-rich (mixture of picture and words). The trial revealed that students performed better on image-rich numeracy problems than on otherwise equivalent word problems (Hoogland, 2016), indicating that students are less hampered by the many difficulties with word problems that are frequently reported (Verschaffel, Greer, & De Corte, 2000; Verschaffel, Greer, Van Dooren, & Mukhopadhyay, 2009). In an experiment in 2013 in the Netherlands the results of this trial were replicated with adult participants. The results are shown in this article and a comparison is made with the results of students from primary and secondary education. It revealed which types of tasks particularly, in both populations, benefitted most from the change from a verbal description of the problem situation, to a mainly depictive description of the problem situation.

## Theoretical perspectives

This study is part of a larger research project to investigate alternatives to the persistent and problematic use of word problems to teach and assess students' ability to deal with numerical problems originating in everyday life. This ability of students is often labeled as numeracy or mathematical literacy, although these concepts have been and are still evolving (Coben, 2003; Geiger et al., 2015; Ruthven, 2016). The sometimes superficial use of the concept is also criticised (Jablonka, 2015).

In current classroom practice, word problems are used predominantly to teach and assess these abilities. Many researchers, however, report serious difficulties in using word problems to assess these abilities (Verschaffel et al., 2000; Verschaffel et al., 2009). The reported difficulties can be related to the steps the problem solver is expected to take to solve the task at hand. Figure 1 shows the diagram used in PISA as a schema for the relevant steps in the problem-solving process. Similar diagrams are used in related research on problem solving and modelling in mathematics education (Blum, Galbraith, Henn, & Niss, 2007; Burkhardt, 2006; Lesh, Post, & Behr, 1987). The reported difficulties seem to appear mainly in the two horizontal steps in the diagram: "formulate the mathematical problem", and "interpret the mathematical results". In the first step (formulate) students are reported to look at these problems with a strong "answer-getting mindset" (Daro, 2013) and a calculational approach (Thompson, Philipp, Thompson, & Boyd, 1994), as if the problem was limited to the right-hand vertical step of the problem-solving process and that solving problems of any kind means getting the "right answer" by conducting a series of operations on the numbers in the problems. In the third step (interpret) students are reported not to take common-sense considerations about the problem into account (Greer, 1997).

We conjectured that the use of images from real life would strengthen the association with real-world situations (Palm, 2009) and therefore decrease the suspension of sense-making (Schoenfeld, 1992) and the strong calculational focus (Thompson et al., 1994). As a paraphrase of the most used definition of word problems (Verschaffel, Depaepe, & Van Dooren, 2014), we suggested the following definition for such image-rich 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”.

Cognitive psychology also offers theories and insights on the effect of depictive and descriptive representations on creativity and problem solving (Schnotz, 2002; Schnotz, Baadte, Müller, & Rasch, 2010; Schnotz & Bannert, 2003). 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. Their line of reasoning is that depictive representations can help students to 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 added to the plausibility of our conjecture, which we tested in our empirical studies and also gives some indications in which kind of problems the effect might be strongest, that is, problems whereof the representation of the problem situation is beneficial for constructing of a (mental) mathematical model needed to solve the problem.

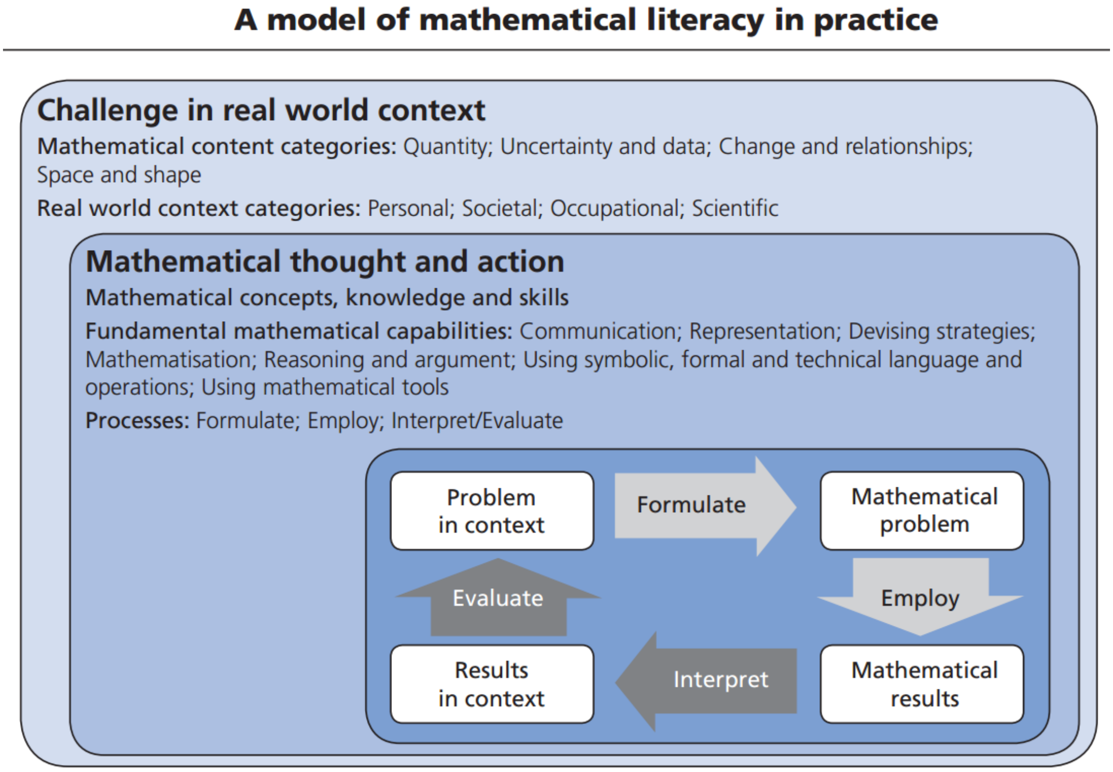

A model of mathematical literacy in practice
Challenge in real world context
Mathematical content categories: Quantity; Uncertainty and data; Change and relationships; Space and shape
Real world context categories: Personal; Societal; Occupational; Scientific
Mathematical thought and action
Mathematical concepts, knowledge and skills
Fundamental mathematical capabilities: Communication; Representation; Devising strategies; Mathematisation; Reasoning and argument; Using symbolic, formal and technical language and operations; Using mathematical tools
Processes: Formulate; Employ; Interpret/Evaluate
Problem in context
Formulate
Mathematical problem
Evaluate
Employ
Results in context
Interpret
Mathematical results

*Figure 1.*A model of mathematical literacy in practice. From OECD (2013a) (p. 26)

## Design of alternatives to word problems

In order to counteract these tendencies and the associated difficulties we designed tasks that were more “authentic” by changing the representation of the problem situation from descriptive to mainly depictive (Hoogland, 2016; Hoogland, Pepin, Bakker, de Koning, & Gravemeijer, 2016). Those tasks were incorporated in an instrument to measure students’ performance on both word problems and image-rich problems in a randomised controlled way. In a trial with students from primary and secondary education our conjecture was confirmed (Hoogland, 2016; Hoogland, Bakker, De Koning, Pepin, & Gravemeijer, submitted). Although the conjecture was confirmed, the results were not straightforward. The students' scores on image-rich problems were slightly higher (2%), which was significant, but with a small effect size (*d* = .09) and the effect of better performance was most noticeable in tasks in the domain of measurement & geometry. The research question for the study reported here is: Does a replication of the original trial with adult participants show the same patterns and results as the original trial with primary and secondary students? In this paper we report on that replication of the original study with adults who participated in the "Groot Nederlands Rekenonderzoek (GNRO)" [Great National Numeracy Survey], a research initiative by the public broadcasting organisations VPRO and NTR, supported by the Netherlands Organisation for Scientific Research (NWO). Individuals of all ages and of all places in the Netherlands could register as participants on the GNRO website and could engage in a series of mathematical tests. We report the results from the trial with students from primary and secondary education in adapted format in the results section for easier comparison.

## The Dutch context

For the international reader, we provide some information on the Dutch educational context. In the Netherlands in 2010 the "Referentiekader Taal en Rekenen" [Literacy and Numeracy Framework (LaNF)] was introduced as a guideline for Literacy and Numeracy education in the age range of 4–18 years (Hoogland & Stelwagen, 2011; Ministerie van OCW, 2009), followed by a very similar version for adult education.

Table 1

*Overview of international frameworks on numeracy and their content domains*

| Framework | Categories | | | | |
|---|---|---|---|---|---|
| TIMSS 2015 – 8th grade | Number | Algebra | Geometry | | Data & Chance |
| PIAAC 2016 | Quantity & Number | | Dimension & Shape | Pattern, Relationship & Change | Data & Chance |
| PISA 2015 | Quantity | | Space & Shape | Change & Relationships | Uncertainty & Data |
| Dutch LaNF 2010 | Numbers | Proportions | Measurement & Geometry | Relations | |

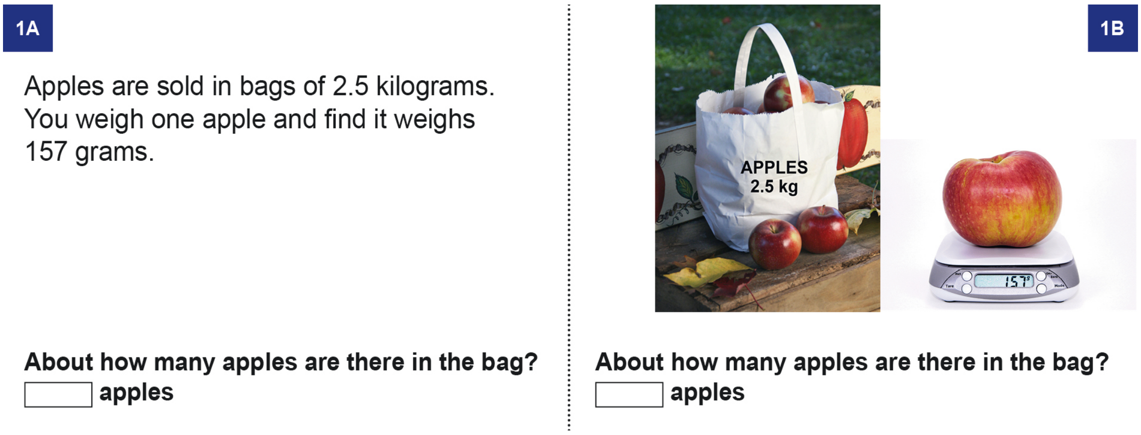
*Note*. Presented by similarity (horizontal).

The content domains in these frameworks resemble the categories used in the international frameworks on numeracy and mathematical literacy, such as TIMSS, PISA and PIAAC (Mullis & Martin, 2013; Organisation for Economic Co-operation and Development (OECD), 2013b; PIAAC Numeracy Expert Group, 2009). Table 1 gives an overview of the content domains used in the various frameworks. It is noteworthy that in the Dutch framework there is more emphasis on proportions, including fractions and percentages, and an absence of focus on uncertainty, chance and data (representation).

## Method

### The instrument

To measure the effect of the change in representation of the problem situation on the performance of participants we used an instrument that was used in both the trial with students in primary and secondary education and in this replication study with adult participants. The trials were held with Dutch language items (Hoogland, 2016); English translations of these items are available under open access (Hoogland& De Koning, 2013). The instrument consisted of 24 items of which 21 items were designed in two versions: word problem and image-rich problem. For every participant a test was composed randomly with 10 or 11 items in each version. The randomly selected items were presented in random order for each participant. In this case a randomised controlled trial was built into the test. Both versions of each item had an equal chance of being selected, independent of any other variables, measured or not.



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**

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

*Figure 2.* An example of an item in two versions: word problem and image-rich problem.

In Figure 2 we show an example of two versions of an item. The items are translated to English for better readability. In the test, each item was presented as a screen-filling problem with an open numerical answer field. The tasks in the research instrument were validated and tested in earlier research activities (Hoogland et al., 2016). The complete set of tasks can be found under open access via the Dutch institute DANS/NWO (Hoogland& De Koning, 2013).

In Table 2 we give an overview of the items in the instrument, evenly distributed across three domains of the LaNF: numbers, proportions, and measurement & geometry. Three tasks in the instrument were in the domain of relations, and were only presented in one version, because of the already visual nature of the items.

Table 2.
*Overview of tasks in the used instrument*

| item | Domain: Numbers | item | Domain: Measurement & Geometry | item | Domain: Proportions |
|---|---|---|---|---|---|
| i04 | TV + DVD | i01 | Apples in bag | i03 | Travel time |
| i05 | Change | i02 | Fuel usage | i06 | Recipe |
| i09 | Money pile | i11 | Double glazing | i07 | Price magazine |
| i12 | Kitchen tiles | i13 | Water bottles | i08 | AEX index |
| i16 | Hamburgers | i14 | Bedroom tiles | i10 | Scale model |
| i17 | Cough syrup | i19 | Cake tin | i15 | Endive |
| i18 | Public debt | i21 | Chocolate boxes | i20 | Winter tires |

## Participants

The research conducted for this paper was a trial with 420 participants from the GNRO research. Table 3 shows the distribution of gender and age categories of these participants. The GNRO was, after registration, an open access public test held in 2013. We cannot consider these participants as a representative sample of Dutch adults. However, we consider the distribution over age and gender diverse enough to draw some tentative conclusions on the results in comparison with the results of the original trial with students from primary and secondary education.

Table 3.
*Number of Participants from GNRO: Age Groups and Gender*

| Age | *n* (%) | Gender | *n* (%) |
|---|---|---|---|
| 15-19 | 15 ( 5.3%) | Male | 115 (40.4%) |
| 20-29 | 61 (21.4%) | Female | 170 (59.6%) |
| 30-39 | 66 (23.2%) | Not stated | 135 |
| 40-49 | 64 (22.5%) | | |
| 50-59 | 31 (10.9%) | | |
| 60-69 | 39 (13.7%) | | |
| 70-79 | 9 (3.2%) | | |
| Not stated | 135 | | |

*Note*. Total sample is 410. *n* is number with percentages taken on stated age and gender in parentheses.

The original trial was conducted in October and November 2011. In that trial 31,842 students from 179 schools geographically spread across the Netherlands, participated. For convenience in comparing we show the results of the participants in the trial with students from primary and secondary education in this section. Table 4 shows the number of participants from the educational streams in the Dutch school system.

Table 4.
*Number of participants in original trial: Age groups and gender*

| Age | *n* | Gender | *n* (%) |
|---|---|---|---|
| 11-12 | 969 (3.1%) | male | 15,310 (49.7%) |
| 12-19 | 30,222 (96,9%) | female | 15,465 (50.3%) |
| Not stated | 680 | not stated | 1,067 |

*Note*. Total sample is 31,842 participants. Age group 11- 12 is primary education, age group 12-19 is secondary education. n is number with percentages taken on stated age and gender in parentheses.

In this original trial we assumed the sample to be representative of Dutch students in the age group 11–19 years.

## Statistical analysis

The statistical analysis focused on the difference in scores on the A-version and the B-version of the 21 paired problems. We conducted a classical analysis using mean, standard deviation, *t*-tests, and Cohen's *d* as effect size to get a general idea of how the separate items contributed to the overall result we found (Cohen, 1988). As a caveat regarding the effect sizes note that we are not dealing with the most common cycle in educational research of measurement – intervention with the participants – measurement. The effect size category lists of Cohen (1988) or Hattie (2009) do not apply to this situation. Changing the representation of the problem situation is not an educational intervention. We are investigating what is the effect on participants' behaviour of such a change and not measuring what they have learned from an intervention or a "treatment".

## Results

We present for both trials the results in the same table format for easier comparison. We compared the results of adults with the results of students from primary and secondary education. We focus in this comparison on the overall test and the results at item level. For the overall result on the test on the data collected in the GNRO we conducted a *t*-test on the mean scores on the A- and B-version items for each participant. We found that the difference in mean was .011 with standard error .001 and *p* = .184 (n.s.).

On item level we conducted a two-sided *t*-test with pooled variances to evaluate whether for each item the scores on the two versions differed significantly. We used a common effect size index, namely Cohen's *d*, for a first general conclusion. The results are shown in Table 5. We found in four paired problems that the scores on the B-version were significantly higher than the scores on the A-versions with effect sizes ranging from .16 to .59. Furthermore, we found in one pair of problems that the scores on the A-version were significantly higher than the scores on the B-version with an effect size of .04. In the 16 other items the differences between the scores were not significant. The results in this replication trial were in most aspects in line with the results in the earlier large-scale student trial, which is discussed in more detail below.

Table 5.
*Results from the GNRO trial, mean and t-test results*

| Item | *N* |  | Mean (SE) |  | *t* -test | effect size *d* |  |
|---|---|---|---|---|---|---|---|
|  | version A | version B | version A | version B | *p* (\|*T*\|>\|*t*\|) | B > A | A > B |
| i1 | 218 | 202 | .899(.020) | .872(.024) | .374 |  |  |
| i2 | 215 | 205 | .823(.026) | .800(.028) | .544 |  |  |
| i3 | 220 | 200 | .836(.025) | .800(.028) | .337 |  |  |
| i4 | 205 | 215 | .951(.015) | .916(.019) | .150 |  |  |
| i5 | 196 | 224 | .898(.022) | .772(.028) | .000 *** |  | .04 |
| i6 | 209 | 211 | .895(.021) | .929(.018) | .218 |  |  |
| i7 | 223 | 197 | .749(.029) | .751(.031) | .955 |  |  |
| i8 | 207 | 213 | .662(.033) | .690(.032) | .537 |  |  |
| i9 | 209 | 211 | .593(.034) | .540(.034) | .274 |  |  |
| i10 | 212 | 208 | .821(.026) | .870(.023) | .162 |  |  |
| i11 | 203 | 217 | .493(.035) | .774(.028) | .000 *** | .59 |  |
| i12 | 212 | 208 | .811(.027) | .789(.028) | .559 |  |  |
| i13 | 202 | 218 | .896(.021) | .858(024) | .233 |  |  |
| i14 | 212 | 208 | .472(.034) | .433(.034) | .423 |  |  |
| i15 | 226 | 194 | .774(.028) | .825(.027) | .198 |  |  |
| i16 | 219 | 201 | .872(.022) | .866(.024) | .845 |  |  |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| i17 | 205 | 215 | .971(.012) | .954(.014) | .355 | | |
| i18 | 194 | 226 | .418(.035) | .540(.033) | .012 ** | .24 | |
| i19 | 203 | 217 | .611(.034) | .691(.031) | .085* | .17 | |
| i20 | 216 | 204 | .533(.034) | .520(.035) | .794 | | |
| i21 | 220 | 200 | .682(.031) | .755(.030) | .096* | .16 | |

*Note*. *N* is number of items tested. Mean is mean score on items (with standard error in parentheses) P(T> t) is result t-test, unpaired, unequal with hypothesis that difference in score is 0; **p*< .10, ***p*< .05,****p*< .01. Cohen's *d* is effect size. Version A is the word problem; Version B is the image rich numeracy problem.

The results of the large scale trial have been published before (Hoogland, 2016). Table 6 highlights only those results that are necessary to make the comparison with the replication sample of this study. For this comparison only we incorporated *p*<.10 as a category – it is not used for further statistical inferences. For the overall results on the test on the data collected in the large-scale school trial, we conducted a *t*-test on the mean scores on the A- and B-version items for each participant. We found that the difference in mean was .019 with standard error .001 and *p*< .001 (***). On item level we conducted a two-sided *t*-test with pooled variances to evaluate whether for each item the scores on the two versions differed significantly. We again used the effect size index, Cohen's *d*, for similar conclusions. The results are shown in Table 6.

Table 6.
*Results for the large-scale school trial, mean and t-test results*

| Item | N | | Mean (SE) | | *t* -test | effect size *d* | |
|---|---|---|---|---|---|---|---|
| | version A | version B | version A | version B | p (\|T\|>\|t\|) | B > A | A > B |
| i1 | 15,878 | 15,964 | .716 (.004) | .720 (.004) | .424 | | |
| i2 | 15,986 | 15,856 | .525 (.004) | .483 (.004) | .000 *** | | .08 |
| i3 | 15,785 | 16,057 | .314(.004) | .290(.004) | .000 *** | | .05 |
| i4 | 15,835 | 16,007 | .826(.003) | .833(.003) | .131 | | |
| i5 | 16,038 | 15,804 | .720(.004) | .828(.003) | .000 *** | .26 | |
| i6 | 15,775 | 16,067 | .631(.004) | .640(.004) | .102 | | |
| i7 | 16,065 | 15,777 | .404(.004) | .416(.004) | .042 ** | .02 | |
| i8 | 16,298 | 15,544 | .303(.004) | .299(.004) | .420 | | |
| i9 | 16,069 | 15,773 | .221(.003) | .213(.003) | .085 * | | .02 |
| i10 | 15,882 | 15,960 | .495(.004) | .525(.004) | .000 *** | .06 | |
| i11 | 15,850 | 15,992 | .145(.003) | .310(.004) | .000 *** | .39 | |
| i12 | 15,871 | 15,971 | .466(.004) | .438(.004) | .000 *** | | .06 |
| i13 | 15,931 | 15,911 | .619(.004) | .641(.004) | .000 *** | .05 | |
| i14 | 15,889 | 15,953 | .040(.002) | .046(.002) | .080 * | .02 | |
| i15 | 15,793 | 16,049 | .394(.004) | .388(.004) | .264 | | |
| i16 | 15,921 | 15,921 | .803(.003) | .815(.003) | .005 *** | .03 | |
| i17 | 15,986 | 15,856 | .803(.003) | .787(.003) | .000 *** | | .04 |
| i18 | 15,847 | 15,995 | .153(.003) | .168(.003) | .000 *** | .04 | |
| i19 | 15,932 | 15,910 | .247(.003) | .284(.004) | .000 *** | .08 | |
| i20 | 15,925 | 15,917 | .130(.003) | .164(.003) | .000 *** | .10 | |
| i21 | 16,044 | 15,798 | .188(.003) | .256(.003) | .000 *** | .16 | |

*Note.* N is number of items tested. M is mean score on items (with standard error in parentheses) P(\|T\|>\|t\|) is result of t-test, unpaired, unequal with hypothesis that difference in score is 0; **p*< .10, ***p*< .05,****p*< .01. Cohen's *d* is effect size. Version A is the word problem; Version B is the image rich numeracy problem.

In the large-scale school trial we found with *p*< .10 in 11 paired problems that the scores on the B-version were significantly higher than the scores on the A-versions with effect sizes ranging from .02 to .39. Furthermore, we found in five paired problems that the scores on the A-versions were significantly higher than the scores on the B-versions with effect sizes ranging from .02 to .08.

## Comparing results

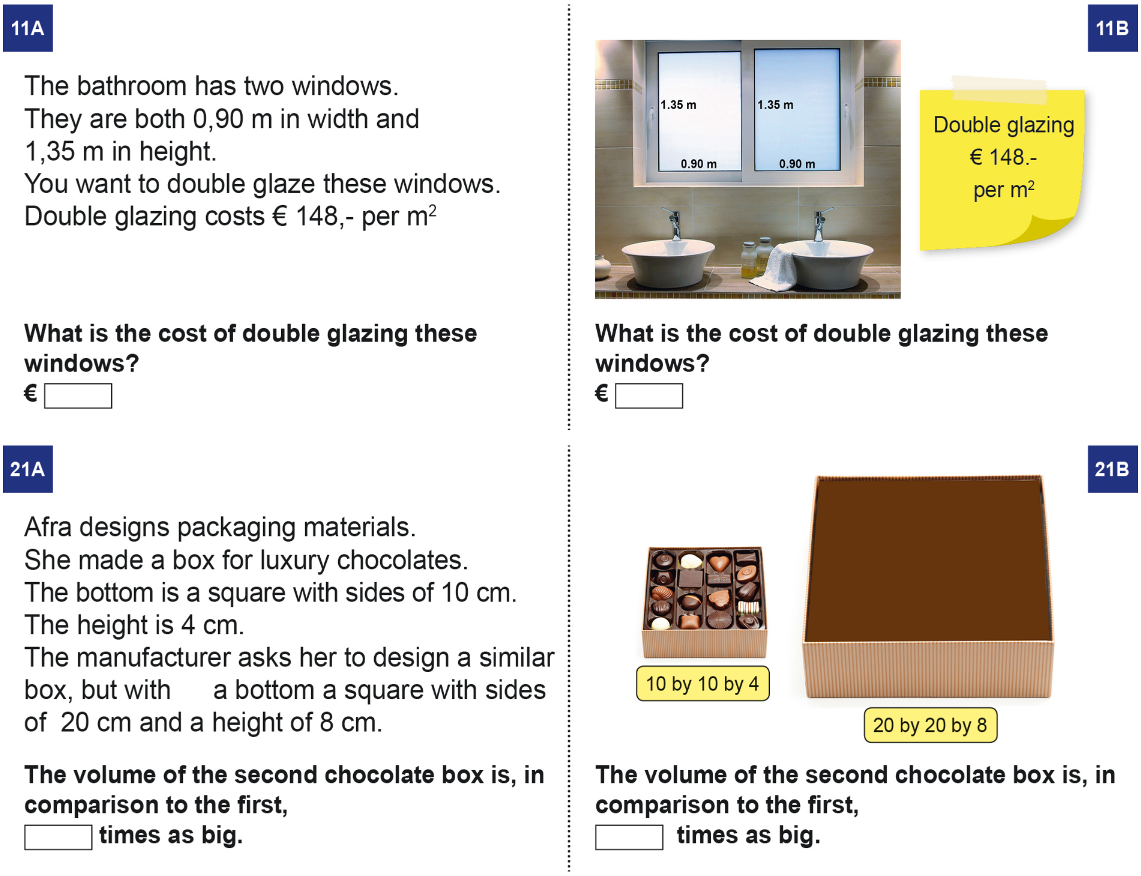
The overall result on performance in this study with adult participants was 1.1 percentage point higher scores on image-rich problems. This was in line with the overall results we found in the large school trial, that is, 1.9 percentage point higher scores on image-rich problems. In almost all items the effect of higher scores on the B-version occurred with a very small effect size, in other items it did not occur. In one item the effect was even opposite. We synthesised the results in Table 7.

Table 7.
*Comparing results of adult and students from primary and secondary education*

| Domain | Population | A > B | A = B | B > A |
|---|---|---|---|---|
| Numbers | Adults | 43% | 14% | 43% |
| | Students | 14% | 71% | 14% |
| Meas. & Geom. | Adults | 0% | 57% | 43% |
| | Students | 14% | 14% | 71% |
| Proportions | Adults | 0% | 100% | 0% |
| | Students | 14% | 43% | 43% |

*Note*. A > B means the results on the word problem version are significantly larger (*p*< .10). A = B means the results are not significantly different (*p*< .10). B>A means the results on the image-rich problem version are significantly larger (*p*< .10).

Solving problems from the domain of measurement & geometry seems to benefit the most from a depictive representation in both populations. For problems in the domain of numbers we see no beneficial effect for either representation, although the deviation is much larger for the adult population. In problems in the domain of proportions only the student population seems to benefit to some extent from depictive representations. We found three tasks in the domain of measurement & geometry that in both trials showed a significant better performance for the image-rich versions. They are shown in Figure 3. This finding corroborates our earlier findings that the change in representation of the problem situation has the greatest positive influence on the performance of the participants in tasks from the domain measurement & geometry. Indeed, in these cases the depictive representation of the problem situation could arguably be beneficial to form a (mental) mathematical model necessary to solve the problem, such as estimating the area in item 11, calculating the content in item 16, and estimating the content in item 21.



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

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

21A

Afra designs packaging materials.
She made a box for luxury chocolates.
The bottom is a square with sides of 10 cm.
The height is 4 cm.
The manufacturer asks her to design a similar box, but with a bottom a square with sides of 20 cm and a height of 8 cm.

**The volume of the second chocolate box is, in comparison to the first,**
□ **times as big.**

21B

**The volume of the second chocolate box is, in comparison to the first,**
□ **times as big.**

*Figure 3.*Three examples from the domain of measurement and geometry with a significantly higher score on the image-rich version.

## Discussion

Assuming the diagram of problem solving in Figure 1 contains essential steps for the solving process (going from the problem situation to the situation model and on to the mathematical model), we argue that the mental activity needed for the necessary steps in the process is interdependent on the mathematics domain of the task. So following the reasoning of Schnotz et al. (2010), in the domain of numbers the mathematical model is primarily computational and thus one dimensional. In that case a mainly depictive representation was presumed not to contribute considerably to the ease with which problem solvers make sense of the situational or mathematical model. In this domain most items gave no significant difference, even one opposite effect. In the domain of proportions the mathematical model is in general more complex than in the domain of numbers, because there is always some activity of (relatively) comparing quantities or comparing a quantity with a whole. A mainly depictive representation was assumed to be beneficial here. At the same time a counter-effect is possible if the mathematical model and the depictive representations are not mutually beneficial, which might lead to an increased complexity experienced by the participant. For tasks from the domain of proportions one could not make a plausible straightforward prediction, whether a mainly depictive representation could help the solvers to construct the appropriate mathematical model and hence help them in solving the problem in a successful way.

In the domain of measurement & geometry the underlying problem situation in itself is two- or three-dimensional. So, a mainly depictive representation of the problem was assumed to help the problem solver to create the appropriate (mental) mathematical model. We saw in both trials that of the four items that significantly favour the image-rich numeracy problem, three are in the domain of measurement & geometry, so this assumption is supported by the data.

In this replication, we found fewer tasks with a significant difference between the A- and the B-versions. With smaller samples, small increases in performance cannot be labelled as statistically significant. Nevertheless the findings give enough incentive for further research in the design of numeracy tasks and the way reality is (re)presented in those tasks.

## Conclusions

Word problems are a dominant feature of both classroom teaching and assessment of numeracy worldwide, and also of large-scale international assessments, like TIMSS, PISA, and PIAAC (Mullis & Martin, 2013; OECD, 2013b; PIAAC Numeracy Expert Group, 2009). Lessons learned from these assessments have been brought together recently, see for instance Tout and Gal (2015). Despite these efforts and despite the dominant use of word problems to teach and assess people's ability to solve practical numerical problems, not much research has been conducted that systematically focuses on the effect on students' performance of changing the verbal representation of the problem situation to a mainly depictive representation or a more authentic representation of the problem situation.

The original trial and this replication have limitations. The participants in the adult sample were not representative of all adults in the Netherlands. And although the replication strengthened some of the conclusions from the earlier large-scale school trial, the conclusions were still based on a limited number of items. More research is necessary to establish whether the results hold for other sets of problems that are paired in the same way as in these trials. The overall difference in results is small and effect sizes related to those results are in most cases very small. Slavin (2016) has recently stated, in a Huffington Post blog "What is a Large Effect Size?", that in educational studies using a randomised controlled trial, effect sizes are seldom found over 0.2, however. The conclusions on the effects of a change in representation of the problem situation can thus be labelled as tentative. At the same time, the results of the change were significant and consistent, and not influenced by other variables, so there is, arguably, enough justification to speak of a small but robust effect.

Our suggestion is that in the task design of future assessments the representation of the problem situation should be taken into account as a factor when interpreting the results.

## Acknowledgements

The data of the replication study described in this paper were obtained in the 'Groot Nationaal Onderzoek' (Great National Survey), a research initiative by the public broadcasting companies VPRO and NTR, supported by the Netherlands organization for Scientific Research (NWO).

The trial replicated in this study, was in 2010 and 2011 funded by the Dutch Ministry of Education under the program Onderwijsbewijs II (Evidence-based Research in Education), under grant number ODB10068.

## References

American Institutes for Research. (2006). *A review of the literature in adult numeracy: Research and conceptual issues*. Washington DC.

Blum, W., Galbraith, P. L., Henn, H.-W., & Niss, M. (Eds.). (2007). *Modelling and applications in mathematics education – the 14th ICMI study*. New York, NY: Springer International.

Burkhardt, H. (2006). Modelling in mathematics classrooms: Reflections on past developments and the future. *ZDM – Mathematics Education, 38*(2), 178-195. doi:10.1007/BF02655888

Coben, D. (2003). *Adult numeracy: review of research and related literature*. Retrieved from London, UK: www.nrdc.org.uk

Cohen, J. (1988). *Statistical power analysis for the behavorial sciences (second edition)*. Hillsdale, NJ: Lawrence Erlbaum.

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

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.

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

Greer, B. (1997). Modelling reality in mathematics classrooms: The case of word problems. *Learning and Instruction, 7*(4), 293-307. doi: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.

Hoogland, K. (2006). Mind and gesture: The numeracy of a vocational student. In M. Horne & B. Marr (Eds.), *Connecting voices in adult mathematics and numeracy: Practitioners, researchers and learners. Proceedings of the adults learning mathematics 12th annual international conference (ALM)* (pp. 150-158). Melbourne, Australia: ACU.

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

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

Hoogland, K., & Pepin, B. (in press). The numeracy of vocational students: Exploring the nature of the mathematics used in daily life and work *Proceedings of the 13th International Congress on Mathematical Education* Hamburg, Germany.

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

Hoogland, K., & Stelwagen, R. (2012). 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.

Jablonka, E. (2015). The evolvement of numeracy and mathematical literacy curricula and the construction of hierarchies of numerate or mathematically literate subjects. *ZDM – Mathematics Education, 47*(4), 599-609. doi:10.1007/s11858-015-0691-6

Lesh, R., Post, T., & Behr, M. (1987). Representations and translations among representations in mathematics learning and problem solving. In C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics*. Hillsdale, NJ: Lawrence Erlbaum.

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

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

OECD. (2013b). *PISA 2015 draft mathematics framework*. 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.

PIAAC Numeracy Expert Group. (2009). *PIAAC numeracy: A conceptual framework*. Retrieved from Paris, France: http://dx.doi.org/10.1787/220337421165

Ruthven, K. (2016). Numeracy in, across and beyond the school curriculum. In D. Wyse, L. Hayward, & J. Pandya (Eds.), *The SAGE handbook of curriculum, pedagogy and assessment* (pp. 638-654). Thousand Oaks, CA.

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

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.

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

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 from http://www.huffingtonpost.com/robert-e-slavin/what-is-a-large-effect-si_b_9426372.html

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.

Tout, D., & Gal, I. (2015). Perspectives on numeracy: reflections from international assessments. *ZDM – Mathematics Education, 47*(4), 691-706. doi:10.1007/s11858-015-0672-9

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

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

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