MASTER THESIS 29-06-2016

WORDS, PICTURES OR BOTH?

The influence of the presentation of contextual numeracy problems on student performance in (pre)vocational education

Researcher
Lynn Buschers
l.j.buschers@student.utwente.nl

Supervisors University of Twente
First supervisor:
Dr. Hans Luyten
j.w.luyten@utwente.nl

Second supervisor:
Prof. dr. ir. Jean-Paul Fox
g.j.a.fox@utwente.nl

External supervisor
Bureau ICE
Drs. M. Binsbergen

UNIVERSITEIT TWENTE.

BUREAU ICE

Acknowledgement

This master thesis is the result of my final project for the master 'Educational Science and Technology' at the University of Twente. The project is conducted at the request of Bureau ICE, an independent Dutch provider of tests and exams. This study will provide more insight into the question how numeracy problems should be presented.

First, I want to thank my supervisors of the University of Twente. Hans Luyten, thank you for the feedback during the whole process. Also I want to thank Jean-Paul Fox for the help with my data analysis and the feedback for the last improvements of my thesis. These supervisors further gave me the freedom to develop myself as a researcher during the execution of this project. Secondly, I want to thank Marc Binsbergen from Bureau ICE. He helped me with designing a good numeracy test and provided me with new insights during this research. Also special thanks to Nicole Bonouvrie for checking the numeracy items I designed. Besides that, I want to thank many others from Bureau ICE for helping me with making my own numeracy tests and making it ready for administration. Last but not least, I want to thank my family and friends for supporting me during my master and my research. They encouraged me and ensured I had a nice break once in a while. Especially, I want to thank Elke for all the feedback she gave me as a critical friend.

I could not have done this without you all.

Lynn Buschers
Enschede, 2016

Summary

In Realistic Mathematics Education (RME), math problems are presented in a rich context in realistic situations. Although it is promising, it also causes new obstacles. A big potential obstacle is the way of presenting the mathematical problems, because this is a crucial part of RME. This is also a problem in the numeracy tests. Because numeracy tests are becoming more important in secondary and vocational education, there is need for an investigation on the presentation of context problems in numeracy tests.

Therefore, this study investigated to what extent the presentation of context problems influences the performance of students of (pre)vocational education on numeracy tests. A distinction was made between numeracy problems in which the numbers are in the text, in an image or both in text and image (repeated information).

First, three variants of a numeracy test were developed, including all three presentations of the context problems. The test was examined with a pilot including a thinking aloud protocol and interviews. Second, an online numeracy test was administered to 301 (pre)vocational students all over the Netherlands. Speed and accuracy were measured and to control for student characteristics a survey was conducted after completion of the test. The results were presented and based on that conclusions were drawn about the effect of presenting of numeracy items. In this case, this effect was very small to negligible, so it can be concluded that presenting redundant information is not as harmful as suggested in the fundamental design theories. While the effects were so small, it is recommended to decide for every individual item which presentation fits the purposes of the numeracy problem.

1. Research description

This section describes the central research problem of this study. First, the context is explained. Second, the issue is viewed from a theoretical perspective and a practical perspective. Subsequently, there is a short description of how the research is conducted. In the last part of the research description, the relevance of this study is elaborated.

1.1. Problem statement

A lot of students in prevocational and vocational education have difficulties with numeracy (Hoogland, Koning & Tanis, 2013). Some researchers claim that the problems can be caused when numeracy problems are offered without a context (e.g. Lave, 1992; Verschaffel, Greer, Van Dooren & Mukhopadhyay, 2009). Therefore, in 1987 Realistic Mathematics Education (RME) was introduced in the Dutch education (Jansen, Van der Schoot & Hemker, 2005). The main goal of RME was providing students with problem situations that are recognizable and perceived by them as useful. In this way, numeracy had to become more relevant to society (Freudenthal, 1973; 1991). Through innovations that are associated with this new development, there is more space for students to actively seek for solutions and express their own way of thinking (Prenger, 2005).

Nowadays realistic numeracy problems are common in a lot of textbooks and tests. Besides that, the Dutch government decided a stronger focus on numeracy in secondary education and vocational education was needed. Consequently, the grade for the numeracy test eventually will be taken into account for the final exam and graduation. This will have serious consequences for students in secondary and vocational education (Vos et al., 2014). Because there is a stronger focus on numeracy, it is important for test developers to know how to correctly present the numeracy problems. The numeracy problems in a test should be designed in a way that they measure the numeracy skills of the student as good as possible. On the one hand students should not be obstructed or distracted by redundant information (Mayer & Moreno, 2003; Van Merrienboer & Sweller, 2005), but on the other hand they should also get enough help to use all the information that is provided.

Based on analysis of test results by Bureau ICE, a Dutch independent provider of tests and exams, practitioners had the impression that students do not get all the required data from images that are presented with the numeracy problem. They presumed that students miss information and are not able to solve the problem, even if they have the numeracy skills to solve it. Therefore, they wondered if it would be better to present the same information in both text and images. Besides that, researchers found that solely using textual presentations could also lead to problems. Here, there is a chance that other students are disadvantaged, because a linguistic task can be more sensitive to interpretation and not only requires numeracy skills, but also language skills (Hoogland et al., 2013). The problem is that these assumptions about giving students more information contradict the fundamental design theories like the cognitive load theory.

In order to remove doubts, it is required that the various options are explored. Then, educational textbooks and tests can be adapted to the needs of the students across the Netherlands. Therefore, this research investigated if there was a difference in the performance on numeracy items with different presentations. This aim was reached by first designing three comparable numeracy tests with three variants of presentation in it, namely the textual variant (LB1), the pictorial variant (LB2) and the variant in which the text and picture contain the same content (LB3). Figure 1 shows an example. Second, the test was administered in (pre)vocational education and the results were analysed. Insight in the performance on comparable items with different presentations resulted in a conclusion about the best way to design realistic numeracy problems and implications for practice.

<table>
<tr><th>Text variant</th><th>Image variant</th><th>Text + image variant</th></tr>
<tr>
<td>Two market stalls are offering strawberries. The first stall sells 500 grams of strawberries for € 1.95. The second stall sells 150 grams of strawberries for € 0.75.

What is the difference in price between stall 1 and stall 2 for a kilo of strawberries?</td>
<td>Two market stalls are offering strawberries.


What is the difference in price between stall 1 and stall 2 for a kilo of strawberries?</td>
<td>Two market stalls are offering strawberries. The first stall sells 500 grams of strawberries for € 1.95. The second stall sells 150 grams of strawberries for € 0.75.


What is the difference in price between stall 1 and stall 2 for a kilo of strawberries?</td>
</tr>
</table>

Figure 1. Example of an item with context adapted into three presentations.

1.2. Relevance

Since the context is decisive in realistic mathematics education, investigation on the way of displaying the numeracy problem is of real importance. Besides that, clarification is needed because literature and practice contradict each other. On the one hand there are a few prominent studies on the design of learning materials (Mayer, 2003; Paas et al., 2003) with the conclusion that information that is already presented in another way should be removed because overload would harm students. On the other hand, practitioners have the impression that students do not get all the required data from images that are presented with the numeracy problem. This research therefore contributes to both theory and practice and reveals if different presentations affect performance and what this means for testing in practice. This may lead to more extensive research on the effects of the presentation of test items within other subjects or for other educational levels. In practice, the presentation of certain numeracy test items can be adjusted to optimize the clarity of the items to positively influence student achievement.

2. Conceptual framework

In this section, the key concepts of this research are elaborated. Since the way of presenting numeracy problems is very important in RME, concepts that are related to this are discussed first. Second, different theories about presenting information are reviewed. Third, the possible consequences of different presentations are discussed. Fourth, other influencing factors that were included in this study are briefly discussed. Finally, the use of response times for measuring differences in presentation is elaborated. Specific literature research was needed to find out if response time was a useful measure in this study. This will be explained in the last section.

2.1. Realistic Mathematics Education

Realistic mathematics education (RME) was developed by educationalists like Freudenthal (1973, 1991), Treffers (1986) and Gravemeijer (1994, 1997). It was the Dutch answer to the urgency to reform teaching and learning numeracy. Connecting numeracy problems with everyday situations was the most important aspect of RME (Hoogland et al., 2013). By making numeracy stronger connected to reality, students develop their own solution strategy and also reflect on that strategy (Willems & Verbeeck, 2011). The term 'realistic' does not mean that the problems are always about everyday life, but the emphasis is on providing students with problem situations that are recognizable and perceived by them as useful (Prenger, 2001). Realistic means that the given problem situation matches the imagination or experience of students (Van den Heuvel-Panhuizen & Drijvers, 2014). The presented problems can be derived from the real world, but this is not necessary as long as the problems are experienced as realistic problems by the students (Van den Heuvel-Panhuizen & Drijvers, 2014). For example, a student is able to imagine a situation in which he or she has to calculate the volume of a spaceship, even if this spaceship does not really exist.

RME already developed in the 1960's and still has considerable impact on the Dutch mathematics education. This RME approach especially had a strong influence on textbooks. Not only the Netherlands, but also other countries, for example the USA and Indonesia are using RME-based mathematics methods (Van den Heuvel-Panhuizen & Drijvers, 2014). Now, several decades after the implementation of RME, researchers are still trying to improve it. More and more, educationalists are trying to increase the authenticity of the problems (Bonotto, 2007). To improve the presentation of context problems in RME, it is important to know how students should deal with them.

Solving context problems consist of three phases (Opmeer, 2005):
1. Translation of the contextual problem to a mathematical problem;
2. Solving this computational problem;
3. The translation of the solution to the context.

This research focused on the first phase of problem solving. According to Opmeer (2005), the first phase, the translation of the contextual problem to a mathematical problem, is always the most difficult phase of a context problem. While this phase makes a context problem more realistic, but also more time-consuming, it remains a topic of discussion. Opmeer (2005), for example, states that in a lot of cases it is better to remove the context and provide plain numeracy problems. This can make a problem easier and draws the focus on the actual calculation. Contrary, Gravemeijer (2006) argues that the contexts are essential for understanding. Educational scientists recognized that knowledge should not be transferred by application of strict rules and procedures, but that pupils should be helped to construct knowledge (Gravemeijer, 2006). Besides that, the learning of rules and definitions makes no sense if students are not capable of applying it to real life situations (Gravemeijer, 2006). So, on the one hand the given context should help students to understand the problem and on the other hand, context problems also help students to learn how to apply calculations in practice. Despite the discussion about whether or not to use context problems, educationalists all agree that these problems need to be designed in a way it meets the needs of students. It is especially important to wisely deal with the decision of how a problem is presented, to overcome issues with overload and wrong interpretations (Opmeer, 2005).

2.2 Design principles

As Opmeer (2005) mentioned before, RME faces the risk that the students continually have to do too many tasks; interpreting the context, calculating and looking for more convenient ways to solve problems. Van de Craats (2007) states that this makes the learning content less integrated in the knowledge of the student. In order to minimize these risks, it is important that the context problems in RME are presented in a way it meets the needs of students. This section addresses two important design principles that apply to the way of presenting context problems, namely the cognitive capacity principle and the dual channel principle.

According to the cognitive load theory (Paas et al., 2003; Sweller, 1999), cognitive capacity for learning can be overloaded, because it is limited. Therefore, most of the time, the information included in an image should not be repeated in additional text. This is to avoid cognitive overload of the short-term memory.

There are three main sources of cognitive load during learning (Van Merrienboer & Sweller, 2005):
1. Intrinsic cognitive load, which depends on the difficulty of the content;
2. Extraneous cognitive load, which depends on the way the information is presented;
3. Germane cognitive load, which is needed for processing, construction and automation of schemas.

The material should be designed in a way that it manages intrinsic load, reduce extraneous load and promotes germane load (Mayer & Moreno, 2003; Van Merrienboer & Sweller, 2005). Due to the fact that the presentation of context problems has no influence on the intrinsic load, only the impact on extraneous and germane load is discussed. Extraneous cognitive load is the amount of cognitive capacity needed for dealing with irrelevant information. This means that the capacity needed for extraneous load reduces the capacity that is left for the understanding of the content (Moreno & Mayer, 2007). One way to do this is removing irrelevant or redundant words or images (Mayer, 2003; 2007). De Jong (2010) agrees with the statement that students should not spend time on irrelevant information, but also discusses some related issues that make this a little more complicated. First, some designs or ways of presentation that cause extraneous load also enhance germane load (De Jong, 2010). Reducing this extraneous load will also have negative consequences for germane load, while germane load has to be promoted (Van Merrienboer & Sweller, 2005). An additional problem is that the distinction between extraneous and germane load depends on the expertise level of the student. Information that is relevant for a beginner can be hindering for an expert (Paas et al., 2003). Additionally, Brünken et al. (2003) mention that cognitive load differences between conditions are related to student characteristics as ability, interest and prior knowledge. This makes it very difficult for designers of contextual numeracy problems to take the needs and characteristics of all kinds of students into account.

It is also unclear if extraneous load really hinders learning. On the one hand, Berends and Van Lieshout (2009) write that because additional information is irrelevant and duplicates the information that is already given, performance can be decreased. On the other hand, other studies indicate that when it is well designed, placed near the information and when it is short, it still can be advantageous (Mayer & Johnson, 2008).

Next to the amount of information that should be presented in a context problem, it is also important to take into account the way the information is processed by students. This leads us to the second principle that that applies to designing context problems; the dual channel principle. The dual channel principle states that students have separate channels for processing verbal and pictorial information (Baddeley, 1999; Paivio 1986). According to Paivio (1986): "Human cognition is unique in that it has become specialized for dealing simultaneously with language and with nonverbal objects and events. Moreover, the language system is peculiar in that it deals directly with linguistic input and output (in the form of speech or writing) while at the same time serving a symbolic function with respect to nonverbal objects, events, and behaviors. Any presentational theory must accommodate this dual functionality." (p 53). Effective use of both channels can very beneficial for processing larger amounts of information. Among many others, Mayer (2003) recommends spreading verbal and pictorial presentations of the same concept in a way that prevents cognitive overload of one channel.

When the information is spread over the verbal and visual channel, one can argue for presenting information in both words and image. This will also fit the needs of different students. Then students can focus on the type of presentation that works the best or use both to be sure they use all the available information. Fiore et al. (2003) state that some argue that performance can be improved, because presenting the material twice results in better reproduction of the information. Others attribute the improvement to the fact that students encode the information both verbally and nonverbally (Fiore et al., 2003). By using multiple ways of representing the same information, the student benefits in the way that each presentation aims on different aspects of the problem (Berends & Van Lieshout, 2008).

2.3. Words or pictures

As stated earlier, designers of numeracy problems often make a choice between using text or images to present the relevant information. In realistic mathematic problems, the amount of language often increases, while the context is explained in words. The linguistic context in realistic numeracy problems has been subject to discussion for as long as these contexts have been around (Verschaffel et al., 2009). Context problems that are purely textual contain several elements that cause obstacles for students (e.g. Abedi &

Lord, 2001; Cummins, Kintsch, Reusser and Weimer, 1988). This makes it more susceptible for different interpretations (Van Eerde, 2009). The consequence is a higher probability of a wrong solutions due to misunderstanding the text. This means that incorrect responses are not always the result of a lack of mathematical understanding (Boonen, 2013; Van Eerde, 2009).

Especially students with specific disorders like dyslexia can experience the negative consequences for the increased amount of language in the context problems. Dyslexia is defined as a disorder whereby the automation of word identification and/or script imaging is incomplete, not developed or severely impeded (Ruijssenaar & Stoop, 1995). As dyslectic students have reading problems, this can lead to lower performance in RME. Because for solving context problems in RME, students must have adequate reading skills, including the necessary vocabulary (Milo & Ruijssenaars, 2003). In addition, good reading comprehension is needed and students have to make a translation from information in the context problem to formal math language (Milo & Ruijssenaars, 2003). Bakker and Beuken (2012) concluded that reading comprehension is the most influential task with respect to solving context problems in RME.

Since a lot of students have difficulties with correctly interpreting textual context problems, more research focuses on the use of visual presentations. Research made clear that visual presentations can make problem solving easier (Boonen, 2013), by making abstract information more concrete (Hamiltion & Rajaram, 2001). In addition, Tall and Thomas (1991) found that images improved the understanding of mathematical higher order concepts. Besides that, Presmeg (1986) showed that visual processing can result in mathematical processing. This can be due to the fact that visual presentations enable students to make connections between their experiences and the concepts in the math problems (Barmby, Bolden, Raine & Thompson, 2013).

However, the presentations should be coherent and complete descriptions of the problem situation. Therefore, the relevant visual and verbal elements have to be connected to each other (Boonen, 2013). If this is not the case, essential illustrations can also decrease accuracy, because understanding the relation between the picture and words becomes difficult (Berends & Van Lieshout, 2008). Besides that, there is no assurance that students recognize the pictorial presentations in the mathematical problems in the way the developer of the mathematical problem has in mind (Barmby et al. 2013).

Although well designed illustrations can be beneficial, it is important to remember that this is not the case for all kinds of illustrations. It depends on the function whether they make problem solving easier or not. According to Van Eerde (2009) illustrations can have three main functions; decoration, support of the context and facilitation of crucial information. This distinction has a lot of overlap with the four functions of illustrations in problem solving; decorative, presentational, organizational and informational (Elia & Philippou, 2004). Bishop (1989) and Elia and Philippou (2004) found that de decorative pictures were not conductive to mathematical problem solving. Presentational pictures show a part or the whole of the content of a mathematical problem (Elia & Philippou, 2004). Organizational pictures focus more on support of the problem solving (Bishop, 1989). Informational pictures provide information that is necessary to solve the problem. This means that the problem is derived from the picture (Bishop, 1989). Besides that, Elia and Philippou (2004) suggest that the way the picture is used, depends on the relationship between the picture and task on the one hand, and on the mental abilities of the student on the other hand. Examples of these mental abilities are spatial and visualization abilities, like the competence to recall, generate, choose and appropriately apply the visualization (Bishop, 1989). Also students' preferences influence the way students apply the information in pictures (Bishop, 1989).

According to literature discussed above, it is evident that context problems have to be designed very carefully. Text and illustrations have to be aligned with the purpose of the numeracy problems and the description of the situation should be a coherent whole. Further, there are no clear indications that there is a large difference in performance between the textual variant and the visual variant. However, based on the design theories mentioned above, a combination of pictures and words can lead to poorer performance. It should be noted that this contradicts the expectations of test developers. Due to these mixed results and advices, the effect of all the three different ways of presenting contextual problems on the performance of the students is investigated.

To measure this in a correct manner, it is important to include individual differences between students in this research, because it appears that these differences can be very decisive in the relation between presentation of context problems and the actual performance. Therefore, the influence of several student characteristics is investigated in a literature study. This is elaborated in the next section.

2.4. Test related, educational, personal and demographical factors

A literature review was conducted to find out what is already known about the influence of other factors on the relation between way of presentation and the performance of the students. The factors can be divided into four different groups: test-related factors, educational factors, personal factors and demographics. All factors are discussed in the next section.

The first group of factors is test-related and consists of the factors domain and test variant. Numeracy domains and variants of the test are factors of the first group. The domain determines the type of numeracy problem. Four numeracy domains can be discerned: numbers, proportions/ratio, measurement and geometry and relations/connections. Students can score differently on the different numeracy domains (Meelissen & Drenth, 2008; Meelissen, Netten, Drent, Punter, Droop & Verhoeven, 2012). It could be that a special variant is more beneficial. For example, the variant with only images can more suitable for problems about measurement and geometry than for ratios. Therefore, it was expected that domain was one of the most influential background variables in this study. The other factor in this group is test variant. The tests were designed in a way that the content of the tests was the same and the different presentations of items were equally divided over the three test variants. For this reason, a relation between the test variant and the performance indicators speed and accuracy is not expected.

The second group of factors consists of study level and type of education and year of study. It is expected that education level has a positive influence on test performance. This is expected, because students in higher levels are on average more intelligent and therefore there performance on numeracy will probably better. Type of education can also influence the performance of the student. For example, students in technical education are better in doing calculations (College van Examens, 2013). Students in technical education score higher than students in the fields healthcare and welfare or economy and services. Besides their higher level on numeracy and mathematics, this can also be explained by their higher motivation for numeracy and mathematics (College van Examens, 2013). Year of study was the last educational factor that was taken into account. Student performance can differ, because when the amount of study years goes up, students have had more learning experiences and their performance on the same test will be better.

The personal factors, preference for a specific variant of the items, learning disorders like dyslexia or dyscalculia and the numeracy skills of the students, belong to the third group of factors that were included in this study. Preference for a specific variant was expected to be one of the most influential factors. It can be related to the performance for different variants, because if students prefer a specific variant, their performance on that variant becomes better (Van Nuland, Dusseldorp, Martens & Boekaerts, 2010). As stated earlier in this conceptual framework, learning disorders can have an effect on the performance of students. How information is presented can influence how well students understand the information. Students with dyslexia can for example have more difficulties and need more time for a text variant of an item compared to a variant with only images. It is also expected that students with dyscalculia are performing lower on the test, because students with dyscalculia have difficulties with numeracy problems in general (Van Luit & Ruijssenaars, 2004). The last factor in this group is numeracy skills. It is obvious that when students have more numeracy skills, they perform better on numeracy tests. Therefore, a strong effect on performance is expected.

The last group of factors that was used was demographics and consisted of gender and age. Research of PISA shows that boys are better in numeracy than girls (Belfi, Levels & Van der Velden, 2015). Therefore it can be expected there is a difference in gender and that male students get higher scores than female students. Furthermore, age may also play a role. For older students it often has been longer ago that they intensively practiced numeracy. In primary school it is practiced almost daily. According to this difference, there were no direct conclusions drawn from literature, therefore there are no expectations regarding this relation.

As we can see, students vary in many ways and this can affect their performance on the numeracy tests. Therefore it is important that this study also investigates if these factors influence the relation between the way of presenting the numeracy problems and the performance of the students. The next section will discuss how performance is measured.

2.5. Student performance

In this study, the performance of students on the numeracy items was measured with the scores and response time, which are operationalisations of accuracy and speed. Both variables can indicate the ability of a student.

Scores (wrong/right) is a very familiar measure, but response times were difficult to measure in the past without use of computers. As a result, it is still relatively unknown how their results should be interpreted. Therefore, the use of response times will be discussed in the next section.

Response time is the time that a student spends on a single item in a test (Lee & Chen, 2011). Many years ago, Thorndike et al. (1926) already came up with a theory about the relation between speed and ability which states that "other things being equal, if intellect A can do at each level the same number of tasks as intellect B, but in a less time, intellect A is better." (p. 33). This means that speed tells something about ability, and that lower response time relates to better performance, if the response stays the same. Dodonova and Dodonov (2013) concluded that the theory of Thorndike et al (1926) holds for baseline performance. They found that, for baseline performance, the high-ability students were faster than low-ability students. This is due to the fact that the more time the student has, the more and better information is available for choosing a good answer (Van der Linden, 2009).

Now computer-based testing has made it very easy to capture the time students spend on individual items (Wise & Kong, 2005), the value of response times has become a point of discussion again. According to Molenaar (2015), response times can be seen in two ways; as a result of an underlying response process, or as an additional source of the latent variable. As a response process indicator, response time is seen as an indicator of the time to start, develop and end the process (Davison, Semmes, Huang & Close, 2012; Molenaar, 2015). In the other case, the speed of performance provides information about intelligence, (Sheppard & Vernon, 2008), educational success (Steinborn, Flehmig, Westhoff, & Langner, 2008), and professional competence (Thompson, Yang, & Chauvin, 2009). These statements are in line with the conclusions of Dodonova and Dodonov (2013), Van der Linden (2009) and Thorndike et al. (1926).

Researchers like Jeon (2015) and Van der Linden (2009) emphasize that ability and speed cannot be separated, because they interact with each other. A number of psychometric efforts have been undertaken to measure intelligence by dealing with the interplay between ability and speed (Jeon, 2015). However, it remains difficult to draw conclusions about how response times can be used as information next to the information that can be derived from the response (Davison, Semmes, Huang & Close, 2012). So nowadays, measurable indicators like item responses and questionnaires are the most common measures that are used to analyse almost every test (Molenaar, Tuerlinckx & Van der Maas, 2015).

It is assumed that the additional information from response times cannot be neglected. Therefore, the relation between the item response and response time is evaluated. Nonetheless, with previous research in mind, this study will analyse response times with some caution.

To sum up, there are no clear hypotheses on the expected outcomes of this research. All three variants have their pros and cons. The advantage of presenting the information both in text and illustration is that the student can decide what to use. A possible danger is that this can cause too much extraneous load. To avoid this overload, one of the other variants can be chosen. When the textual variant is chosen, there is a risk that students with reading problems fail, while they are able to do the math. On the other hand, the visual variant can also give problems when it comes to interpreting the context problem. This study reveals if some advantages and limitations of the variants result in differences in student performance. The performance was measured using the binary responses and response times.

3. Research design and methods

In this section, the research questions with the corresponding research model are presented. Next, the research design for answering the questions is presented. Other parts of this section give a more detailed insight in how the research is conducted. Here details about the respondents are provided, used instruments are explained and the procedure is described. This section concludes with information about the data analysis.

3.1. Research question and model

The research question is an evaluative question used to improve the presentation of numeracy test items. This study is conducted in both prevocational and vocational education. The question if the presentation influences the performance can be answered by measuring the response and response time per numeracy item, because they both provide information about the performance of the student. This is also shown in the research model in Figure 2.

The following question is central to this research: *To what extent can the presentation of a context problem influence the performance of students of (pre)vocational education on numeracy tests?*

To answer the overarching question, the research question is divided in two sub-questions;

1. To what extent can the presentation of a context problem influence the response (wrong/right) of students of prevocational and vocational education on numeracy tests?
2. To what extent can the presentation of a context problem influence the response time of students of prevocational and vocational education on numeracy tests?

Based on the expectations from the literature study, the influence of various background variables (educational level, type of education, preferences, age, gender, suffering from dyslexia or dyscalculia, numeracy skills, test variants and numeracy domains) was also examined.

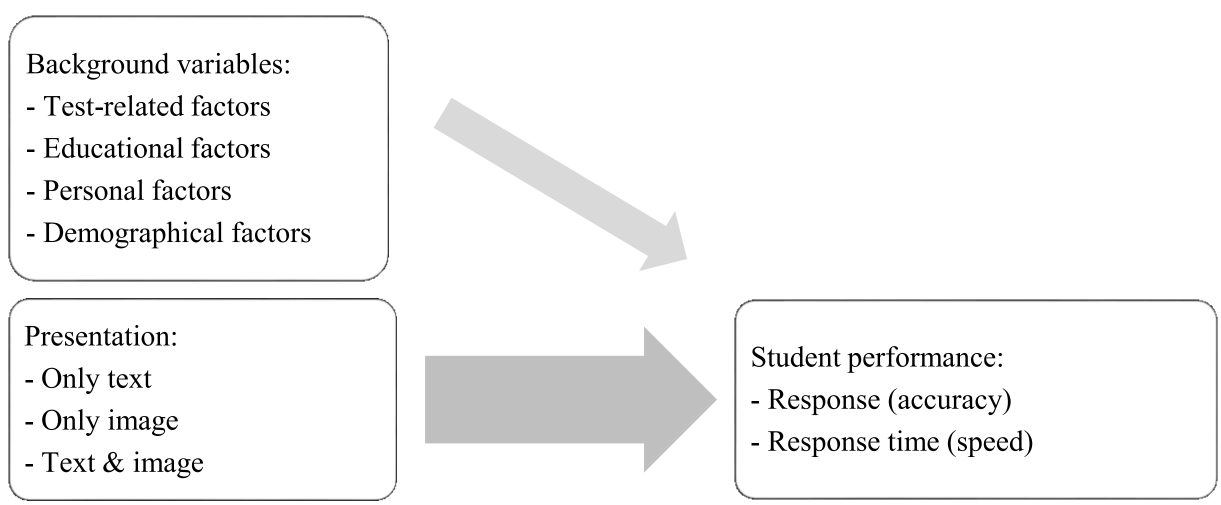


Figure 2. Research model.

3.2. Research design

In Figure 3 the research design is presented. The first part of the research, the literature study and the pilot with thinking aloud protocol and interviews, was qualitative. The main part of the study was focused on the quantitative measurement of the effects of different variants of presentation. At the end of the numeracy test, there was a short questionnaire. The qualitative data from literature and the thinking aloud protocol was used to maximize the appropriateness of the quantitative instruments, the numeracy tests.

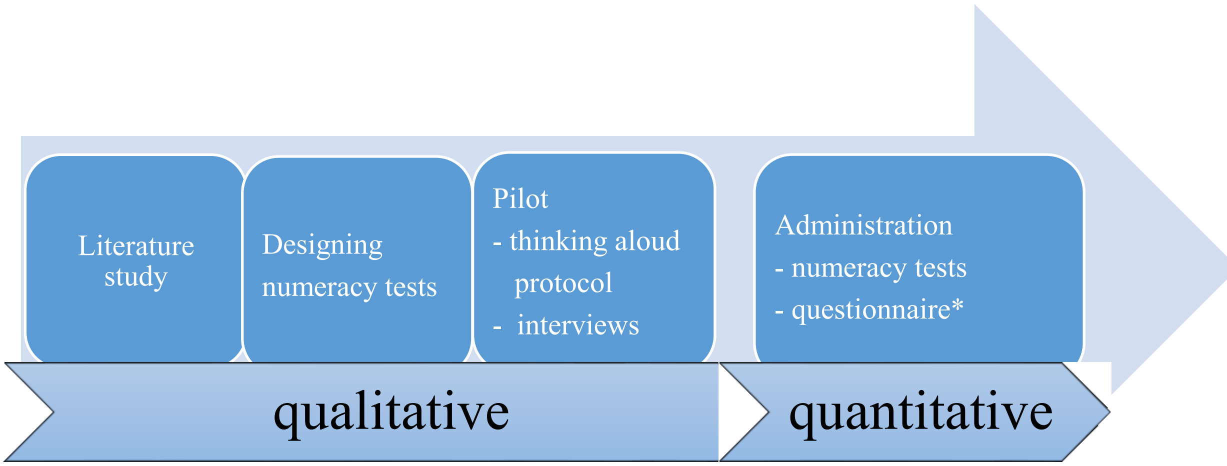


Figure 3. Research design. * There is one qualitative question about preference in the questionnaire.

The quantitative part was a randomized experimental research on the effect of the presentation of context problems in numeracy tests on student performance. In the experiment students were randomly assigned to one of the three tests. In every test, one of the presentation variants (LB1, LB2, LB3) of the same item was presented. Every test contained all three presentations, the textual variant, pictorial variant and the variant in which both the written description and the picture are displayed, so the preferences of the individual students did not affect the outcomes. The outcomes are the response time per item and on the responses per item. In this way the study provided objective data. The analysis of the data gave insight in the influence of the presentation of context problems in numeracy. This research design also gave the opportunity to generalize the conclusions to comparable groups of students.

3.3. Respondents

The respondents in this research were students from the three levels of prevocational education and the third level of vocational education. Both prevocational education and vocational education are included, because the national numeracy test will be compulsory for prevocational and vocational education in the near future and these students have to master the same level, namely 2F.

In total 301 students participated in the study. The participated schools were all Dutch schools from different cities, with different religious or educational visions and different student populations. More than half of the students, namely 58.6%, were from secondary schools, of which 15.1% participated in basic prevocational programmes (vmbo basisgerichte leerweg), 28.4% in advanced prevocational programmes (vmbo kadergerichte leerweg) and 15.1% mixed/theoretical programmes (vmbo gemengde/theoretische leerweg). The other 41.4% of the students followed studies on the third level of vocational education (mbo niveau 3). Within these tracks all five sectors are represented; Services (23.7%), Nature (14.5%), Trading (18.4%), Technics (16.9%) and Healthcare (26.6%).

To give an overview of the demographic student characteristics that were analysed, descriptive statistics for gender and age were also explored. In this study 44.7% of the students were male and 55.3% was female. Of all students, 88.9% of the students were between 14 and 18 years old. The youngest student was 12 years old and the oldest was 21 years old. As stated in the conceptual framework, learning disorders can have an effect on the performance of the student. Therefore, this is also taken into account in this research. Of the participating students 11.2% had dyslexia, 4.7% had dyscalculia and 1.7% had both dyslexia and dyscalculia. The majority of the students had no related learning disorder (82.4%). Besides that, the students also indicated which of the item variants they preferred the most. The most popular variant according to the questionnaire, which was administered after the test, was the variant with both text and an image (43.2%). The second favourite was the variant with only an image (23.6%) and the least favourite was the variant with only text (6.0%). A very small amount of students stated that they did not care how the test items were presented (2.6%). The remaining students did not answer the question (24.6%).

The students were randomly assigned to one of the three variants of the numeracy test and are almost equally divided over variant 1, variant 2 and variant 3[1], respectively 34.2%, 32.9% and 32.9%. In the numeracy test, items of different domains were presented. Most of the items were related to the domains Measurement and geometry (37.1%) and Ratios (37.1%) The other items were of the domains Numbers (14.3%), Relations (5.7%), a combination of Measurement and geometry and Ratios (2.9%) and a combination of Measurement and geometry and Relations (2.9%). As explained in the section about the design of the numeracy test, not all kinds of questions of the domains could be translated to the presentations used in this study. Therefore, not all domains are equally represented.

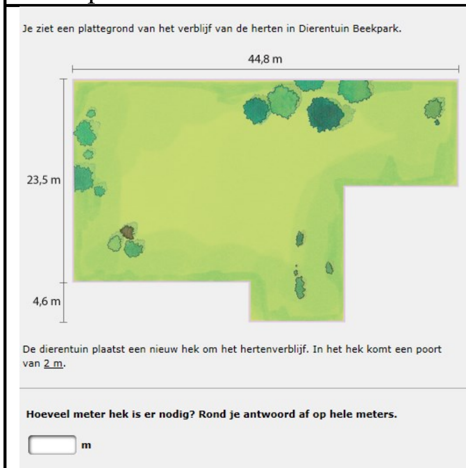
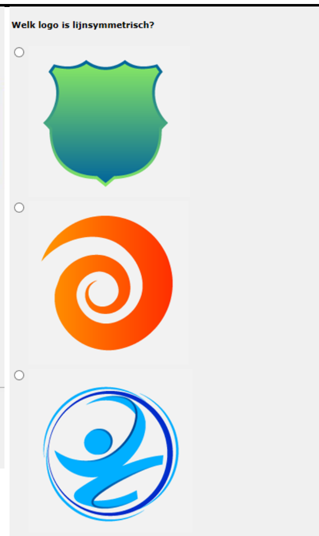
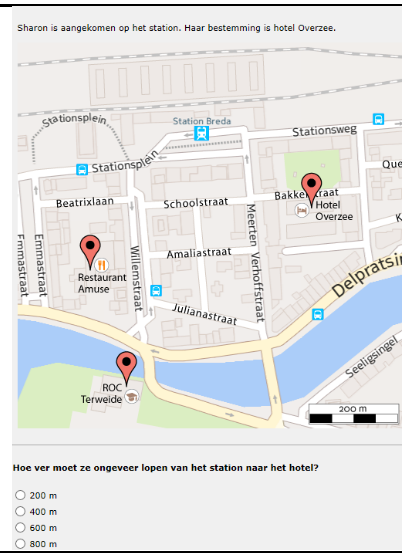
3.4. Instrumentation

Numeracy test

Three numeracy tests were developed with each 35 context items. The tests comprised equal numbers of items that contained only written text, information presented in an image or both written text and an image that present the same information. The content was the same for every respondent. The reference level of the items was all the same, namely 2F. The test items were in Dutch, because the respondents that made the test were from schools in The Netherlands. The test was presented in an online environment. In this way, the response time could be measured and the answers were also automatically checked by the system. This made it possible to find out if the presentation of the context problems had influence on the responses and time students needed for problems with a specific presentation.

The first step in the design process was selecting numeracy items. All context items in the item bank from Bureau ICE were scanned on usability in the test. For some items it was not possible to create different variants, because they contained graphs, tables, pictures or descriptions that were not suitable for adaptation. For items with tables and graphs it is not desirable to adapt it, because interpreting the graph or table is the most important ability that the student has to show. For similar reasons, it is also not useful to adapt items with symmetry[2]. Figure 4 gives an idea of the type of items that were not suitable and therefore were excluded from the test.

[1] The difference between the test variants is explained in Table 1, which shows the test designs.
[2] Symmetry is a mathematical concept about the shape of objects. The second image of Figure 4 illustrates the problem with adapting numeracy problems about symmetry.

Example items		
		
Reason for exclusion		
The total circumference of the area has to be calculated. Describing the way the field looks and the length of the given sides is too difficult. It is obvious that an item like this should not be translated into a textual variant.	This item is about symmetry, the images are needed to test if the students know how to solve this problem. To test this numeracy skill, the images are necessary.	A map is also to complicated to describe in words. Besides that, students have to measure the lengt of the streets in the image.

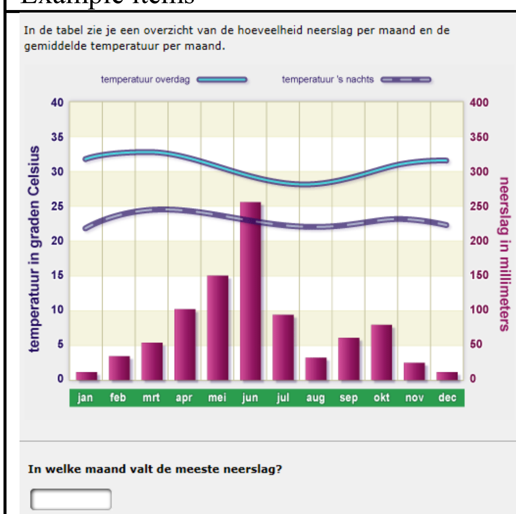
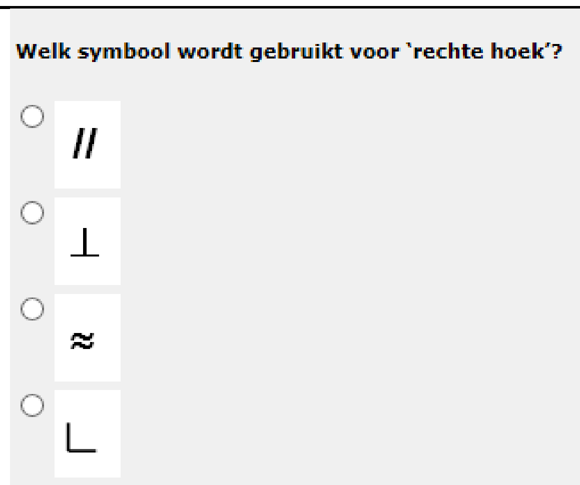
Example items	
	
Reason for exclusion	
This numeracy skill cannot be tested without the graph. Therefore, it is not desirable to translate the information into text.	This item is about symbols. Students have to learn this, so it is not possible to change the way the problem is representated.

Figure 4. Items that were not usable.

For the items that seemed suitable for the numeracy test, the other two variants were made. The original items were not adapted, because this item was already checked by several experts of Bureau ICE. This means that it met the standards that Bureau ICE had set according to the checklist (see Appendix A).

Eventually, from 49 items, three variants were constructed; the textual variant LB1, the pictorial variant LB2 and the variant with text and picture LB3. Figure 5 shows an example of the three variants for an item in the test.

Variant LB1	Variant LB2	Variant LB3
The box is 12 centimetres long, 5 centimetres wide and 4 centimetres high. In the box there is space for three chocolates per $60cm^3$ How many chocolates fit in the box?	Ruben buys a box of chocolates for his mother.  In the box there is space for three chocolates per $60cm^3$ How many chocolates fit in the box?	The box is 12 centimetres long, 5 centimetres wide and 4 centimetres high.  In the box there is space for three chocolates per $60cm^3$ How many chocolates fit in the box?

Figure 5. Difference between the three variants LB1, LB2 and LB3.

Most of the items had one information component that was translated to a picture, but for some items two or three components were converted to a picture. This was done to find out if there also was a difference for the amount of information that was translated to text or images. All pictures that were constructed for this research were adapted versions of images without copyrights or restrictions or self-designed images. The other images were already attached to the items in the item bank and these are especially designed for the tests of Bureau ICE. When all items were finished, they were checked by two experts from Bureau ICE. They have analysed all the variants of items, in order to see if they met the guidelines (as described in Appendix A) and whether the content of the two constructed variants was equal to the content of the original items. Three test variants were made, all containing one of the variants of every context item. The amount of items with information in the text (LB1), information in an image (LB2) or both (LB3) was divided equally (see Table 1).

Table 1

Test design

Test variant 1	Test variant 2	Test variant 3
Item 1 - LB1	Item 1 – LB2	Item 1 – LB3
Item 2 – LB2	Item 2 – LB3	Item 2 – LB1
Item 3 – LB3	Item 3 – LB1	Item 3 – LB2
Item 4 – LB1	Item 4 – LB2	Item 4 – LB3
Etc…	Etc…	Etc…

After the context items were placed in the tests, also the introduction texts, example items and the informed consent form were inserted. The first page was about informed consent. Here the researcher explained the set-up, the anonymity, the possibility to withdraw from participation without a reason. By continuing to the next page, the student agreed with the conditions. Next, information was given about the amount of items that had to be answered and the use of a calculator. After the example items, there was an announcement that the test started. The three numeracy tests each contained 35 context items.

Pilots

The pilot indicated if specific items had to be added or removed. This was done with a thinking aloud procedure and interviews. For the thinking aloud procedure, there was no instrument developed. The researcher just listened to the respondent and asked additional questions, when the respondent was not giving enough detailed information. The researcher also used a recorder and made notes. This is elaborated further in the procedure section and the data analysis section.

For the interview an interview scheme was developed. This scheme is presented in Appendix B, and was written in Dutch. The interview is semi-structured and contains questions about student characteristics and the quality and usability of the numeracy tests. There are some questions about student characteristics; educational level and prior/previous education, age, grade for numeracy, dyslexia or dyscalculia and gender. An example of an item about personal characteristics is; *'Do you have dyslexia or dyscalculia?'* Further, questions about the quality and usability of the numeracy test were for example; *'Did you see words in the test that you did not understand? If so, which words?'*.

Based on results of the thinking aloud procedure and the interviews the numeracy test was improved. The survey that followed after the numeracy test was very similar to the content of the interview. The pilot indicated that it was not needed to add or remove specific items.

3.5. Procedure

To investigate if the adapted items of the numeracy test were suitable to use, a pilot was conducted in two ways. First, after the procedure was explained and permission was asked for recording, five respondents filled out the tests with a thinking aloud protocol. Then the participant was asked to talk aloud, while solving a problem. This request was repeated if necessary during the problem-solving process. The participant did not give an interpretation of his or her thoughts; he or she just mentioned the thoughts as they come to mind. After the procedure, the researcher and the respondent shortly discussed the procedure and the use of the data. Second, another five respondents filled out the test without interruption and were interviewed after the test. The participants got the same information as the first group of participants beforehand and afterwards. The second group of the pilot was also recorded with permission.

Both ways of data collection were used, because with a thinking aloud procedure students do not have an overall view on the test, because they focus on their thought about specific items. Contrary, this is the case for respondents who answered all items without telling what they see and think. In this way, additional information is obtained. The pilot revealed if it was necessary to adapt the items or remove some items from the test. This improved the focus of the survey that was conducted after the numeracy test.

For conducting the numeracy tests, several schools in the Netherlands were approached. When schools chose to participate in the study, the school management and the teachers were informed in advance and the researcher decided the moment of testing together with the teachers involved. The students received data in order to log in for the online test and survey. Before the start of the test, the researcher explained the meaning and procedures of the study. It took about an hour to fill out the numeracy test including the questionnaire, but this varied from student to student.

The numeracy test and survey contained informed consent. The informed consent form is presented in Appendix B (in Dutch). This means that, before the test started, all respondents are informed about the goals and method of the survey, the estimated time to complete it and guaranteed anonymity. The conditions about informed consent had to be accepted to get access to the test and survey. The quality of the instruments and the procedures were guaranteed by the Ethics Commission of University of Twente, which indicates that the study was executed according to the rules and norms of University of Twente.

When students finished the test and the survey, they got their score on the test. This score had no consequences and it was not used for grading. Afterwards, the researcher explained how the data were used and there was room for general questions about the study.

3.6. Data analysis

The data from the interviews and thinking aloud protocol used in the pilot was recorded and during the pilot notes were made. The transcripts of the interviews and thinking aloud protocol were coded with Atlas.ti. There was no fixed coding scheme beforehand, so the coding was inductive. When a pattern was observed, the approach became more deductive. Then, the data was analyzed again for specific themes and remarks.

To investigate differences in performance between the three presentations (the only text presentation, the pictorial presentation and the combined presentation), effects on responses and response times were measured. Before these effects could be analyzed, it was important to control for the average working speed and the ability of the student. These person characteristics influence, together with item characteristics, the responses and response times on the numeracy test. This is shown in the following equations.

Response = Person (ability) + Item + Error (+ variant)
Response time = Person (speed) + Item + Error (+ variant)

In these equations, 'variant' is placed between brackets, because it is not clear if this variable has an effect on the response and response time. This study will answer this question.

The cirt program of Fox, Klein Entink, and van der Linden (2007) was used to compute an ability and a working speed score for each student given his/her responses and response times. The cirt program is based on the two-parameter item response theory (IRT) model to measure ability and a log-normal response time model to measure working speed, while accounting for the correlation between ability and speed. Missing data were assumed to be missing at random, and responses and response times were imputed under the model to deal with the incomplete cases. Therefore, for each student an ability measurement and working speed measurement was computed, despite missing values.

After the ability and speed of the students were determined, the analyses were conducted with the program SPSS Statistics 23. To find out if there was a difference between the scores on the test for the three presentations (the only text presentation, the pictorial presentation and the combined presentation), a Wald test was conducted. The F test was used for the analyses on the response times. This allows the researcher to make the overall comparison on whether means differ. The values were calculated, considering $\alpha = 0.05$. If significant differences were revealed, post hoc analyses were conducted to see on which variant students scored higher or had answered it faster. The influence of the four groups of factors (test related, educational, personal and demographical) was also checked by adding them to the model.

The validity of the designed numeracy items was ensured by the by two experts from Bureau ICE. They have analysed all the variants of items, in order to see if they met the guidelines (as described in the checklist) and whether the content of the constructed variants was equal to the content of the original items. The reliability of the three variants of the numeracy test also appeared to be good enough. Values around .8 are good (Field, 2009). Cronbach's alpha for variant 1, variant 2 and variant 3 was respectively .92, .92 and .91.

Further, the assumptions of normality were assessed using the values for skewness and kurtosis. The dependent variables are normally distributed when the values are around zero (Field, 2009). Besides that, the histograms and the P-P Plots for the dependent variables were checked. The histogram and P-P Plot also showed that the assumptions for normality were not met. Therefore, the natural logarithm of the response times was calculated and used in the analyses. This also reduces the effect of possible outliers.

Homogeneity of variance can be assessed using Levene's test (Levene, 1960), which tests if the variances in different groups are equal. According to Field (2009) it does not really matter if this assumption is met, because when the group sizes are equal, the assumption can be ignored. This means that Levene's test does not apply to this research.

4. Results

This section the results obtained from the pilot and the numeracy test with the questionnaire are presented. The pilot was used to improve the numeracy test and questionnaire. The results of the numeracy tests and the questionnaire were used to find out if there was a difference between the three ways of presenting the numeracy items. Besides the main effects, interaction with influencing factors was also investigated. Outcomes of all analyses are described below.

4.1. Pilot

For the pilot three students from prevocational education and six students from vocational education were selected by purposeful and convenience sampling, because the pilot interviews ask for motivated and involved respondents that can give constructive feedback (Dooley, 2001; Onwuegbuzie & Leech, 2007).

The thinking aloud procedure showed how the students solved the numeracy problems. The elements that were unclear to the participants were not caused by mistakes or aspects of the presentation. The unclear elements were parts of numeracy problems, so adaptation of the items was not desirable. For example, there was a misunderstanding about the image. The image showed the radius of a round pillow and in the question the diameter was asked. Actually recognizing the radius in the image and then convert it to the diameter is a numeracy skill that has to be mastered. In the thinking aloud procedure, it became also clear that the participants found it very hard to calculate what the relationship between two different volumes was, when they had a textual variant (variant LB1) of an item. For example, a few participants were struggling with calculating how many magazines fit into a box. They tried several times and in the end they often drew an image themselves. These problems were not visible when the participants had to solve a variant of the item where an image was displayed (variant LB2 and LB3).

The thinking aloud protocol did show that for one item it was not possible to fill in the answer box. That had to be fixed after the pilot. Besides that, the images were almost always too large. The participants had to scroll to first use the calculator and then re-read the question and fill in an answer. This was confusing and took extra time. This was also mentioned in all the interviews. Therefore, the images were adapted in a way that the information, the question, the answer and the calculator were simultaneously visible without scrolling.

The interviews also revealed that the respondents preferred the variant with both text and image (LB3) over the other ones. With this variant it was possible to check information twice. Sometimes the respondents looked only at the image and often they were more focused on the image than on the text, but overall they were not affected by redundant information in the text. However, one participant did mention that she found it confusing when there was shown redundant text. She preferred the variant with only an image, because this was the most clear.

Further, all respondents claimed there were no difficult words in the numeracy test or questionnaire, so nothing was changed there. The total length of the test was also good. The time the participants spend on the test varied from 25 minutes to 50 minutes when they made the test without thinking aloud. With thinking aloud, the participants needed from 50 to 70 minutes. They mentioned that they took more time to describe in detail what they saw and thought. The time participants spent on the test also varied because there was no time pressure and differences in numeracy skills, but all of them indicated that the test did not have to be longer or shorter.

The interview scheme of the pilot interviews was adequate. The same scheme was used for the questionnaires in the numeracy tests, except for the last section with the comments for improvement of the test.

4.2. Numeracy test

The results of the numeracy tests and questionnaire were used to analyse the effect of the presentation of the numeracy items on the performance of the students. Table 2 shows the means, standard errors and confidence intervals of the responses for the text variant, image variant and the variant with both text and image. The means are the average scores on the items. When a student gave an incorrect answer he or she got the score 0 and when a student answered the item correctly he or she got the score 1.

Table 2

Means, standard error and confidence intervals of the responses

			95% Wald Confidence Interval	
	M	SE	Lower	Upper
Text variant (n = 2700)	.33	.01	.31	.35
Image variant (n = 2711)	.34	.01	.32	.36
Text + image variant (n = 2704)	.36	.01	.34	.38

*M=mean, SE=standard error.

A Wald test was used to measure the effect of the predictors. Here the item characteristics, numeracy skills, the independent variable variant and the dependent variable response were entered in the model. The Wald test showed that the difference in responses between the text variant (n = 2700, M = .33, SE = .012), the image variant (n = 2711 M = .34, SE = .012) and the variant with both text and image (n = 2704, M = .36, SD = .012) were not statistically significant, $X^2(\mathrm{N} = 8115) = 3.73$, $p > 0.05$. This means that the presentation of a numeracy problem did not influence the score on the items. This is shown in Table 3.

Table 3

Wald test for the effect of variant on the (binairy) response

	Type III		
	X^2	*df*	*p*
Variant	3.73	2	.16
Item	1399.50	34	.00*
Numeracy skills	1428.07	1	.00*

* Effect is significant at the 0.055 level (Bonferroni correction).

To measure if background variables influenced the relation between the presentation of numeracy items and the responses, a second Wald test was used. Here the background variables (domain, test variant, preference, learning disorder, numeracy skills, educational level, study year, type of education, age and gender) were also entered in the model. When item and domain are entered separately in the model, the influence of domain cannot be measured, because domain is an item characteristics. Therefore, the interaction effect between item and domain is used instead of the item variable. In this way, it was possible to add domain to the model. The output is shown in Table 4. From this analysis, it can be concluded that domain had a significant effect on the responses, $X^2(\mathrm{N} = 5326) = 339.88$, $p < 0.05$. This means that some domains are more difficult than others. The other background variables had no significant influence on the responses on the numeracy items. The effect of the variants on response was again not significant, $X^2(\mathrm{N} = 5326) = 1.89$, $p > 0.05$.

Table 4

Wald test for the effect of variant and background variables on the (binairy) response

	Type III		
	X^2	*df*	*p*
Numeracy skills	714.22	1	.00*
Item * Domain	727.32	29	.00*
Domain	339.88	5	.00*
Variant	1.89	2	.39
Test variant	.00	1	1.00
Preference	.68	1	.41
Dyslexia/Dyscalculia	.31	1	.58
Level of education	.13	1	.72
Study year	.29	1	.59
Type of education	.69	1	.41
Age	.02	1	.89
Gender	.01	1	.93

* Effect is significant at the 0.055 level (Bonferroni correction).

Domain appeared to have a significant effect on the responses. Therefore, it was investigated whether there was an interaction effect between domain and variant. Only significant effects were entered into the model. In this way, a parsimonious model was created. In Table 5 is shown that there is no interaction between domain and variant.

Table 5

Wald test for the effect of variant and domain on the (binairy) response

	Type III		
	X^2	*df*	*P*
Numeracy skills	1421.25	1	.00*
Domain	508.89	5	.00*
Item * Domain	1072.94	29	.00*
Variant	1.28	2	.53
Variant * Domain	8.30	10	.60
Item * Variant * Domain	67.61	58	.18

* Effect is significant at the 0.055 level (Bonferroni correction).

To see if the variants, contrary to the results for the binary responses, influenced the response times, means for the different variants were compared with an F test. In the table below (Table 6) the means, standard errors and confidence intervals are presented in seconds. It should be noted that this is only an indication, because the scores were not corrected for possible outliers. Therefore, the natural logarithms of these values were used in the conducted analyses. This corrects possible outliers.

Table 6

Means, standard error and confidence intervals of the response times in seconds

	M	SE	95% Wald Confidence Interval Lower	Upper
Text variant (n = 2701)	36.34	1.02	35.27	37.45
Image variant (n = 2704)	37.60	1.02	36.49	38.74
Text + image variant (n = 2704)	38.17	1.02	38.17	40.53

*M=mean, SE=standard error.

Next, an F test was conducted to investigate if there was a relation between the presentation of the numeracy items and the response times.[3] Here the item characteristics, speed, the independent variable variant and the dependent variable response time were entered in the model. The test showed that the difference in response times between the text variant (n = 2701, M = 3.59, SE = .02), the image variant (n = 2704 M = 3.63, SE = .02) and the variant with both text and image (n = 2704, M = 3.67, SE = .02) were statistically significant, $F(2,8109) = 7.08$, $p < .05$. with b = -.079. The model is presented in Table 7. When the effect is converted into seconds, it turns out that on average the text variant (LB1) is answered .92 seconds faster than the version with text and image (LB3).

Table 7

F test for the effect of variant on the response time

	F	*Df*	*Sig.*
Corrected Model	23.09	333	.00*
Intercept	105103.21	1	.00*
Variant	7.08	2	.00*
Item	39.07	34	.00*
Speed	19.20	297	.00*

* Effect is significant at the 0.055 level (Bonferroni correction).

Post hoc tests using the Bonferroni correction (Table 8) showed that the response times of the text variant were lower than the response times of the variant with both text and image (M = 3.63, SD = 1.13 versus M = 3.68, SD = 1.09, respectively), which was statistically significant ($\beta = .0809$, $p < .05$). However, there was no significant difference between the text variant compared to the image variant and the image variant compared to the variant with both text and image.

[3] In the analyses, the natural logarithm of the response times was used, therefore means, standard errors and effects differ from the numbers in Table 7.

Table 8

Comparisons of response times for the three variants

					95% Confidence Interval	
(I) Variant	(J) Variant	*MD* (I-J)	*SE*	*p*	Lower	Upper
Text	image	- .04	.022	.20	-.09	.01
	text and image	- .09*	.022	.00	-.13	-.03
Image	text	.04	.022	.20	-.01	.09
	text and image	- .04	.022	.17	-.09	.01
text and image	text	.09*	.022	.00	.03	.13
	image	.04	.022	.17	-.01	.09

Note. MD = Mean Difference
*Effect is significant at the 0.055 level (Bonferroni correction).

To measure if background variables influenced the relation between the presentation of numeracy items and the response time, a second F test was used. Here the item characteristics, speed, dependent variable response time, the independent variable variant and the background variables (domain, test variant, preference, learning disorder, numeracy skills, educational level, type of education, age and gender) entered in the model. As in earlier analyses, instead of item, the interaction between domain and item was used in the model. The output is shown in Table 9. From this analysis, it can be concluded that domain was the only background variable that had a significant influence on the response time, $F(5,5321) = 7.50$, $p < .05$. The effect of the variants on response time stayed significant, $F(2, 5321) = 3.76$, $p < .05$.

Table 9

F test for the effect of variant and background variables on the response time

	F	*df*	*p*
Speed	2011.82	1	.00*
Variant	3.09	2	.05*
Item * Domain	31.82	29	.00*
Domain	7.50	5	.00*
Test variant	.01	1	.91
Preference	.00	1	.97
Dyslexia/Dyscalculia	.38	1	.54
Level of education	2.12	1	.15
Study year	.04	1	.85
Type of education	.11	1	.74
Age	.00	1	.97
Gender	.03	1	.87

*Effect is significant at the 0.055 level (Bonferroni correction).

While domain had a significant effect on the response time, an additional analysis was conducted to find out whether domain influenced the main effect. This is presented in Table 10. The model shows that the effect of variant on the response time stayed significant, $F(2, 5321) = 3.60$, $p < .05$. There is no interaction between domain and variant. This means that differences between variants arenot domain specific.

Table 10

F test for the effect of variant and domain on the response time

	F	*df*	*p*
Speed	5389.10	1	.00*
Domain	8.15	5	.00*
Domain * Item	52.06	32	.00*
Domain * Item * Variant	1.47	58	.01*
Variant	3.60	2	.03*
Domain * Variant	1.29	10	.23

*Effect is significant at the 0.055 level (Bonferroni correction).

Besides the objective results in terms of student performance, research was also conducted on the preferences of the students. According to the questionnaire which was administered among the students and the conversations that took place after taking the test, it can be concluded that the vast majority of students had a preference for the variant which contained both text and images (LB3). One of the participants mentioned; 'In this way you get a clearer picture of the situation, because you get more information’. This quote was representative for what students thought about this way of presenting numeracy items.

5. Discussion

In the beginning of this thesis, it was stated that research on the presentation of numeracy problems is important, because literature and practice gave different conclusions. Therefore, this study investigated if there was a difference in three presentations of the contextual numeracy problems (LB1, LB2, LB3) by administering numeracy tests in (pre)vocational education. Scores and response times on the numeracy items were used to see if there was a difference in performance for the three variants. This section elaborates on the conclusions of this research, by answering the research questions based on the results. Finally, this section concludes with the implications for theory and practice, by giving advice about further research and about how to use these results in test practice.

5.1. Effect of presentation on performance

The aim of this study was to investigate whether there was a difference in the way in which questions are presented in numeracy tests for (pre)vocational education.

The research question that was central to this research was: *To what extent can the presentation of a context problem influence the performance of students of (pre)vocational education on numeracy tests?*

To answer the overarching question, the research question was divided in two sub-questions;

1. To what extent can the presentation of a context problem influence the response (wrong/right) of students of prevocational and vocational education on numeracy tests?
2. To what extent can the presentation of a context problem influence the response time of students of prevocational and vocational education on numeracy tests?

To answer these questions, multiple analyses were conducted. The analyses showed that there was no significant difference between scores for the different ways of presentation (LB1, LB2, LB3). Contrary, there was found a small effect for the response times. Here, the text variant (LB1) proved to take less time than the variant with both text and image (LB3). The average difference between the text variant and the variant with both text and image was less than a second. This effect is thus small, but did not disappear when background factors were added in the model. There was no clear difference found between the text variant and image variant (LB1 vs. LB2). There was also no difference between the image variant and the variant with both text and image (LB2 vs. LB3).

When taking a closer look at the average scores on the items, it seems that the version with both text and image (LB3) results in the highest scores. However, a statistical significance effect could not be shown with the conducted analyses. Also adding the background variables did not change that.

This means that based on this information the research questions can be answered. The first sub-questions about how much the presentation of a context problem would influence the response, can be answered with the conclusion that there was no detectable difference in influence on the responses. Students scored equally well on the different types of presentations of the numeracy items.

The second sub-questions relates to the impact of the presentation of a context problem on the response time. This question can be answered with the conclusion that there was a relatively small influence, because the mean difference was less than a second. Here, the text variant of the numeracy problems was answered slightly faster than the variant in which both text and the image were displayed. Besides the main effects, this study also took possible background variables into account. For domain it was further investigated if one of the three ways of presenting information was more suitable for the one than for the other. Although the domains differed in difficulty, there were no indications that a specific domain would benefit more from presenting information in one of the three variants than other domains. This means that the effects in this study were only visible on the item level, which makes it hard to give a general advice according to the presentation of items.

The overarching question can now be answered, combining the conclusions of both sub-questions. The presentation of contextual numeracy problems has almost no effect on the performance of students of (pre)vocational education. It can be concluded that, for this population it does not really matter which variant is chosen. If only significant effects are taken into account, it can be said that the variants with only text are more favourable over the others, because it was the type of item that costs the least amount of time, while the scores on text items were not significantly lower. While the difference between the variants was so small, it is doubtful whether it is worth to take this into account, especially when there is no really high time pressure.

Therefore, it can be concluded that the influence of the presentation of numeracy items is so small that test developers do not have to consider changing their items on this point. The conclusions of this study are compared to the expectations based on the literature. Presenting both information in both text and image was not as harmful as the major design theories claimed it was. In fact, it was not harmful at all. On the other hand, potential benefits of providing information in both text and images at the same time, were not found either. With these contradicting assumptions, it was no surprise that if differences were found, they were small. In some cases, one variant is more suitable and in other cases other variants suit better. Nonetheless, it is remarkable that students in this study had a strong preference for the variant in which both text and images were presented, while this had no effect on their performance.

Altogether, looking at the results of this study, it is advisable to pick a presentation that best fits the content of the specific numeracy problem. Unnecessary information or decoration should be avoided, but one way of presenting information is not per definition better or worse than the other. The analyses showed that differences were found on the item level. This makes it very difficult to draw conclusion about what is the best way for presenting a specific item. There were for example no differences found across domains, which means that differences are on the item level. It is not clear why for some items it makes a little difference which variant is used and for other items it does not matter. It should be noted that this only applies to the presentation of relevant information. Therefore, the results can differ when context problems in numeracy tests contain more decorative images or more (difficult) words than necessary.

5.2. Implications for theory and practice

Although there is a lot of variation in the sample and the research design is strong, it is still hard to say if the results are generalizable over other target groups. For numeracy tests on reference level 2F, results for students from (pre)vocational education will probably be the same. A lot of the students are not reaching the expected level for numeracy and therefore find the numeracy test very difficult. This can lead to smaller differences in the results, because the scores are low and the response time can be too short. This choice was not made beforehand, because the numeracy tests for (pre)vocational education are, as stated in the problem statement, becoming more and more important. Therefore, to know if results can be generalized, research can be conducted with a test that is easier for the aimed population. It can be the case that when tests are exactly on the level of the students, differences between variants become more visible. When the test is peaked around the student's true ability, higher measurement accuracy can be achieved (Kröhne & Martens, 2011). So called borderline students, who have just enough numeracy skills to answer the items, give the most information about the performance. These students are namely on the line between just able to solve the numeracy problem and not being able to solve the numeracy problem (Mortaz Hejri & Jalili, 2014). Contrary, when numeracy items become too easy, the presentation also can become even less important. This can be investigated in future research.

Difficulties with the numeracy problems have probably negatively influenced the motivation. Besides that, during observations of the administration of the tests, it became clear that most students were not always motivated due to other causes. First of all, the students were not motivated for the numeracy tests, because there were no consequences. They knew that they would not get a grade. Research shows that grades are a valid reward for several reasons. It is a direct linear function of the task performance (Pulfrey, Darnon & Butera, 2012). In this study, students only got an indication in the sense of the percentage correct. Second, it provides the students with diagnostic information. Third, it has a social symbolic function. This is a significant predictor in the relation between reward and motivation (Pulfrey, Darnon & Butera, 2012). Next to the motivational problems related to the conducted numeracy tests, students are not motivated for learning numeracy in general. The national numeracy test has no consequences for graduation yet. Therefore, students do not pay attention during lectures and do not put a lot of effort in practicing, even if there level is far below the aimed reference level they should reach.

These reasons ensure that the numeracy tests were low-stakes assessments and that it is unclear how much effort the students put in. Without enough effort, performance is likely to suffer, which results in scores that are lower than can expected based on the abilities of the students (Wise and Kong, 2005). It is in fact well known that talent guarantees success when students are lazy (Kuhn & Ranger, 2015). It is the interaction between the willingness to achieve and the abilities of the student that determines the actual performance on the test (Kuhn & Ranger, 2015). When students are not motivated, it makes validity of the resulting test scores more complicated (Wise & Kong, 2005).

Besides the fact that it influences the scores on the test, motivation can also affect the response time. On the one hand, some response times were very low, because students did not want to give too much effort in solving the problems. On the other hand, some response times were very high because students were distracted while solving the numeracy problems. Therefore, it is recommended to ensure high motivation of the students in further research on this topic. This can be done in two ways, by selecting students that are intrinsically motivated, or by motivating selected students. In the first option, students are motivated, because they like solving numeracy problems as an activity. In the second option, students are motivated, because of the external prods, pressures or rewards (Ryan & Deci, 2000). When the numeracy test becomes a high stakes test, this will influence the willingness and most students will perform at their maximum ability (Kuhn & Ranger, 2015). Another way to increase the external motivation is giving a reward for a good performance, for example by raffling an iPad.

Another reason for the fact that no significant effects on the responses were found, could be caused by the way of assessment of the numeracy items. In this study, answers were automatically categorized as wrong or right answers. A wrong answer resulted in the score 0 and a right answer in the score 1. A disadvantage is that there is no distinction between no response and a response that was almost completely correct. This can be done by giving a specific amount of points to each question, depending on the number of correct steps in the calculation. The assessment is more difficult in online testing, because there are a lot of options for getting to the correct answer. Therefore, in further research on this topic, other ways of assessment have to be used like paper-and-pencil tests or more advanced online testing programs.

Research on paper-and-pencil tests (PPT) are not only interesting, because of the scoring options. It can also reveal if these tests have the same results as the results in this study, using computer-based testing (CBT). According to Hanel, Goldhammer, Naumann and Kröhne (2016) ICT changed the way information is presented and received by the test takers, which can affect the comprehension and therefore also their performance. This means that the conclusions of this study cannot automatically be generalized to PTT. For example, Pomplun, Frey and Becker (2002) concluded that reading on a computer screen is slower compared to PPT.

Besides that, within-item navigation, amount of items per page and computer skills can also influence response behaviour of students. First, within-item navigation (scrolling) can cause difficulties when the information and question cannot be presented on one page (Kingston, 2009). This was also confirmed in the pilot for the improvement of the numeracy tests in this study. Participants indicated that it was annoying that all the information was not visible at one glance. To ensure that the entire item fitted on a page, the size of the images was adjusted. Second, space on the computer screen is one of the practical reasons that a CBT often presents one item at the time. Although this is also feasible with PPT, most of the time there are more items on a printed page (Schwarz et al., 2003). This can make the two types of testing less comparable (Bennet, 2003). Third, comparability of performance of students with different levels of computer experience can also be an issue. One of the differences can be the method for entering text (Clariane & Wallace, 2002), like a calculation. In this study, students also mentioned that they were not used to an onscreen calculator. Therefore, it is suspected that some students had more difficulties answering the questions, because they had less experience with online testing.

Based on the points mentioned above, it can be concluded that it is important to take into account of the characteristics of computer based tests. This applies to both the generalizability to other online numeracy tests as to testing the effects of various ways of presenting information in paper-and-pencil tests.

Another recommendation for further research is to conduct a similar study in which the accuracy is measured on a continuous level, instead of on a binary level (wrong/right). By measuring responses on a binary level, no observable effect was found in the way of presenting information on the scores on the numeracy items. Opposed to this, response times were measured on a continuous level and here significant differences were found. So, there is a possibility that small differences in responses (higher scores on one variant) could be found when the responses are also measured on a continuous level.

Besides that, Proctor and Vu (2003) concluded that there is a relationship between a lower working speed and making fewer mistakes. According to the results of this study, it seems that the responses on the variant with both text and image (LB3) are a little better. This outcome is also possible if we look at the effect of the presentation of numeracy items on the response times, because students needed more time for the variant with both text and image. The average scores on the variants of the numeracy problems and the fact that the text and image variant (LB3) takes more time than the text version (LB1), gives reason to suspect that the variant with text and image may lead to better performance. It is recommended to investigate this further, including the above mentioned recommendations regarding the methodology.

Finally, a bigger sample can be used to generate more data and find more significant effects. The test must not be longer. This makes it hard to administer the tests at schools and it also has a negative influence on the concentration of students. Further, the same research design is recommended, to ensure that the results are independent of the participant.

6. Reference list

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219-234.

Baddeley, A. D. (1999). *Essentials of human memory*. Hove: Psychology Press.

Bakker, E. E., & van den Beuken, M. A. (2012). *Contextopgaven: De invloed van begrijpende leesvaardigheid, rekenvaardigheid en rekenangst.* (Master thesis). Available from Utrecht University Repository.

Barmby, P., Bolden, D., Raine, S., & Thompson, L. (2013). Developing the use of diagrammatic presentations in primary mathematics through professional development. *Educational Research*, *55*(3), 263-290.

Belfi, B., Levels, M., & Van der Velden, R. (2015). De jongens tegen de meisjes. *ROA Reports*, *5*. 1-149.

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning*. Princeton: Educational Testing Service.

Berends, I. E., & van Lieshout, E. C. D. M. (2009). The effect of illustrations in arithmetic problem-solving: Effects of increased cognitive load. *Learning and Instruction, 19*, 345–353.

Bishop, A. (1989). Review of research on visualization in Mathematics Education. *Focus on Learning Problems in Mathematics, 11*(1), 7-15.

Bonotto, C. (2007). How to replace word problems with activities of realistic mathematical modelling. *Modelling and applications in mathematics education, xx,* 185-192.

Boonen, A. J. H., Van der Schoot, M., Van Wesel, F., De Vries, M. H., and Jolles, J. (2013). What underlies successful word problem solving? A path analysis in sixth grade students. *Contemporary Educational Psychology, 38*, 271–279.

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*, 53–62.

Clariane, R., & Wallace, P. (2002). Paper-based versus computer-bases assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33,* 593–602.

College van Examens (2013). *Rapportage invoering centrale toetsing en examinering referentieniveaus Nederlandse taal en rekenen*. Utrecht: College van Examens.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive psychology*, *20*(4), 405-438.

Davison, M. L., Semmes, R., Huang, L., & Close, C. N. (2012). On the reliability and validity of a numerical reasoning speed dimension derived from response times collected in computerized testing. *Educational and Psychological Measurement*, *72*(2), 245-263.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science, 38*(2), 105-134.

Elia, I., & Philippou, G. (2004). The Functions of Pictures in Problem Solving. *International Group for the Psychology of Mathematics Education*, *2*, 327-334.

Dodonova, Y. A., & Dodonov, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence*, *41*(1), 1-10.

Dooley, K. (2001). *Social research methods.* Upper Saddle River, NJ: Prentice Hall.

Eggen, T.J.H.M. & Sanders, P.F. (1993). *Psychometrie in de praktijk*. Arnhem: Cito.

Field, A. (2009). *Discovering statistics using SPSS*. London: Sage publications.

Field, A. P., Miles, J. N. V., & Field, Z. C. (2012*). Discovering statistics using R: and sex and drugs and rock 'n' roll*. London: Sage publications.

Fiore, S. M., Cuevas, H. M., & Oser, R. L. (2003). A picture is worth a thousand connections: the facilitative effects of diagrams on mental model development and task performance. *Computers in Human Behavior, 19*(2), 185–199.

Fox, J.-P., Klein Entink, R.H., van der Linden, W.J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software, 20*(7), 1- 14.

Freudenthal, H. (1973). *Mathematics as an Educational Task.* Dordrecht: D. Reidel.

Freudenthal, H. (1991). *Revisiting Mathematics Education*. Dordrecht: Kluwer Academic Publishers.

Gravemeijer, K.P.E. (1994). Educational development and developmental research in mathematics education. *Journal for Research in Mathematics Education*, 443-471.

Gravemeijer, K.P.E. (1997). Solving word problems: a case of modelling? *Learning and Instruction*, *7*(4), 389-397.

Gravemeijer, K.P.E. (2006). Wiskunde leren is complexer dan je denkt: Reactie op vraagtekens bij realistisch reken-wiskundeonderwijs. *Volgens Bartjens Ontwikkeling en Onderzoek, 25*(4), 33-36.

Hahnel, C., Goldhammer, F., Naumann, J., & Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Computers in Human Behavior*, *55*, 486-500.

Hamilton, M., & Rajaram, S. (2001). The concreteness effect in implicit and explicit memory tests. *Journal of Memory and Language*, *44*(1), 96-117.

Hickendorff, M. (2013). The language factor in elementary mathematics assessments:
computational skills and applied problem solving in a multidimensional IRT framework. *Applied Measurement in Education*, *26*(4), 253-278.

Hoogland, K., de Koning, J., & Tanis, O. (2013). *Rekenen in Beeld*. Utrecht: APS.

Janssen, J., Van der Schoot, F., & Hemker, B. (2005). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool. Uitkomsten van de vierde peiling in 2004. Periodieke peiling van het onderwijsniveau. PON-reeks, nr. 32.* Arnhem: Cito.

Jeon, M. (2015). Differential Response Speed: Is it Really a Nuisance? *Measurement: Interdisciplinary Research and Perspectives*, *13*(3-4), 169-172.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22,* 22–37.

Kröhne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitung für Erziehungswissenschaftm 14*(2), 169-186.

Kuhn, J. T., & Ranger, J. (2015). Measuring speed, ability, or motivation: A comment on Goldhammer (2015). *Measurement: Interdisciplinary Research and Perspectives*, *13*(3-4), 173-176.

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: Harvester Wheatsheaf.

Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modelling*, *53*(3), 359-379.

Levene, H. (1960). Robust tests for equality of variances. *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, *2*, 278-292.

Mayer, R. E. (2003). *Learning and instruction*. Upper Saddle River, NJ: Prentice Hall.

Mayer, R. E., & Johnson, C. I. (2008). Revising the redundancy principle in multimedia learning. *Journal of Educational Psychology, 100*, 380–386.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, *38*(1), 43-52.

Meelissen, M. R. M., & Drent, M. (2008). *TIMSS-2007 Nederland: trends in leerprestaties in exacte vakken in het basisonderwijs*. Enschede: Universiteit Twente.

Meelissen, M. R. M., Netten, A., Drent, M., Punter, R. A., Droop, M., & Verhoeven, L. (2012). PIRLS-en TIMSS-2011. *Trends in leerprestaties in lezen, rekenen en natuuronderwijs*. Enschede: Universiteit Twente.

Meijerink, H. P., Letschert, J. F. M., Asselt, R., Bergh, H. H., Craats, J., Haterd, L. A. G. J., ... & Snippe, J. (2008). *Over de drempels met taal en rekenen: Hoofdrapport van de expertgroep doorlopende leerlijnen taal en rekenen.* Enschede: SLO.

Milo, B. F., & Ruijssenaars, A. J. J. M. (2003). Instructie en leerlingkenmerken. *Tijdschrift voor nascholing en onderzoek van het reken-wiskundeonderwijs*, *22*(1), 27-33.

Molenaar, D. (2015). The Value of Response Times in Item Response Modelling. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3-4), 177-181.

Molenaar, D., Tuerlinckx, F., & Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197-219.

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, *19*(3), 309-326.

Mortaz Hejri, S., & Jalili, M. (2014). Standard setting in medical education: fundamental concepts and emerging challenges. *Medical Journal of the Islamic Republic of Iran*, *28*, 34, 1- 6.

Onwuegbuzie, A. J., & Leech, N. L. (2007). Sampling designs in qualitative research: Making the sampling process more public. *The qualitative report*, *12*(2), 238-254.

Opmeer, M.R., (2006). Vraagtekens bij realistisch reken-wiskundeonderwijs. *Volgens Bartjens Ontwikkeling en Onderzoek, 24*(4), 25-28.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: recent developments. *Educational psychologist*, *38*(1), 1-4.

Paivio, A (1986). *Mental presentations: a dual coding approach.* Oxford: Oxford University Press.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62,* 337–354.

Prenger, J. (2001). Vocabulaire hindernissen bij wiskunde. *Toegepaste Taalwetenschap in Artikelen, 66,* 53-68.

Prenger, J. (2005). *Taal telt!: Een onderzoek naar de rol van taalvaardigheid en tekstbegrip in het realistisch wiskundeonderwijs.* Groningen]: Rijksuniversiteit Groningen.

Presmeg, N. C. (1986). Visualisation and mathematical giftedness. *Educational studies in mathematics*, *17*(3), 297-311.

Proctor, R. W., & Vu, K.-P. L. (2003). Action selection. In A. F. Healy & R. W. Proctor (Eds.), Experimental psychology (pp. 293–316). Vol. 4 in I. B. Weiner (Editor-in- Chief) Handbook of psychology. New York: Wiley

Pulfrey, C., Darnon, C., & Butera, F. (2013). Autonomy and task performance: Explaining the impact of grades on intrinsic motivation. *Journal of Educational Psychology*, *105*(1), 39.

Ruijssenaar, A.J.J.M., Stoop, J.W. (1995): *Dyslexie. Afbakening en behandeling*. Den Haag: Gezondheidsraad.

Sheppard, L. D., & Vernon, P. A. (2008). Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences*, *44*(3), 535-551.

Steinborn, M. B., Flehmig, H. C., Westhoff, K., & Langner, R. (2008). Predicting school achievement from self-paced continuous performance: Examining the contributions of response speed, accuracy, and response speed variability. *Psychology Science Quarterly*, *50*(4), 613.

Schwarz, R. D., Rich, C., & Podrabsky, T. (2003, April). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cognitive science*, *12*(2), 257-285.

Sweller, J. (1999). *Instructional Design in Technical Areas.* Melbourne: ACER Press.

Tall, D., & Thomas, M. (1991). Encouraging versatile thinking in algebra using the computer. *Educational Studies in Mathematics*, *22*(2), 125-147.

Thompson, J. J., Yang, T., & Chauvin, S. W. (2009). Pace: An alternative measure of student question response time. *Applied Measurement in Education*, *22*(3), 272-289.

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence.* New York, NY: Teachers College Bureau of Publications.

Treffers, A. (1986). Analyseren en ontwikkelen van reken/wiskundeonderwijs vanuit twee verschillende basisconcepten. *Pedagogische Studiën, 63*(1), 14-26.

Van de Craats, J. (2007). Waarom Daan en Sanne niet kunnen rekenen. *Nieuw Archief voor Wiskunde*, *8*(2), 132-136.

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

Van der Linden, W. J. (2009). Conceptual issues in response-time modelling. *Journal of Educational Measurement*, *46*(3), 247-272.

Van Eerde, D. (2009). Rekenen-wiskunde en taal: een didactisch duo. *Panamapost. Rekenwiskundeonderwijs: onderzoek, ontwikkeling, praktijk*, *28*(3), 19.

Van Luit, J. E. H., & Ruijssenaars, A. J. J. M. (2004). Dyscalculie, zin en onzin. *Rekenwiskundeonderwijs: onderzoek, ontwikkeling, praktijk*, *23*(2), 3-8.

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational psychology review*, *17*(2), 147-177.

Van Nuland, H. J., Dusseldorp, E., Martens, R. L., & Boekaerts, M. (2010). Exploring the motivation jungle: Predicting performance on a novel task by investigating constructs from different motivation perspectives in tandem. *International Journal of Psychology*, *45*(4), 250-259.

Verhelst, N.D., Glas, C.A.W. and Verstralen, H.H.F.M. (1995). *One parameter logistic model (OPLM).* Arnhem: CITO.

Verschaffel, L., Greer, B., Van Dooren, W., & Mukhopadhyay, S. (Eds.). (2009). Words and worlds: Modelling verbal descriptions of situations. Rotterdam/Boston/Taipei: Sense Publishers.

Vos, A., de Schutter, J., Beverdam, A., Zwart, R., Reichardt, R., van Schoonhoven, R., ..., van Nieuwkerk, H. (2014). *Over rekenen gesproken.* Ede: Steunpunt taal en rekenen mbo.

Willems, W. & Verbeeck, K. (2011). *De leraar als regisseur. Opbrengstgericht rekenonderwijs bij de invoering van de referentieniveaus in PO en VO onderwijs.* 's-Hertogenbosch: KPC Groep.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183.

Words, pictures or both?

7. Appendices

Appendix A: Checklist construction & screening numeracy items

Checklist constructie & screening rekenen	
Formele eisen	
Itemformat	☐ Is het juiste itemformat gebruikt (MC, open)? ☐ Is het itemformat juist en volledig ingevuld? ☐ Is het item juist gelabeld?*
Dekking toetsmatrijs	☐ Sluit het item aan bij het referentieniveau, Domain, Domainonderwerp en vraagtype (kaal/context) zoals aangegeven in de toetsmatrijs?
Itemkwaliteit	
Algemeen	☐ Voldoet het item aan de algemene afspraken constructie & screening rekenen (zie volgende pagina's)? ☐ Sluit het item aan bij het beoogde referentieniveau (doelen, aantal rekenstappen, voorbeeldopgaven conceptsyllabus)? ☐ Sluit het item aan bij het beoogde rekendomein en domeinonderwerp? ☐ Is het vraagtype (open/MC) geschikt om het beoogde doel te meten? ☐ Sluit het item aan bij de doelgroep (mbo/vo/po)? ☐ Is het item cultureel, religie en sekse neutraal (zie ook richtlijnen PC)? ☐ Is er voldoende variatie tussen items op toetsniveau (bijv. niet meerdere items met 25%/¼/een kwart)?
Tekst	☐ Is de tekst functioneel voor het beantwoorden van het item (geen overbodige tekst)? ☐ Is de tekst eenduidig te interpreteren? ☐ Is de tekst kort, helder, eenvoudig en taalkundig juist geformuleerd (woordgebruik, zinsbouw, actieve formulering)?
Afbeelding	☐ Voldoet de afbeelding aan de voorwaarden zoals beschreven in het document 'Checklist screening assets'? ☐ Is de afbeelding functioneel voor het beantwoorden van het item (nodig om vraag te kunnen beantwoorden of ter vervanging van context)? ☐ Is de afbeelding helder en eenduidig beschreven in het format assets? ☐ Staat de afbeelding op een logische plaats ten opzichte van de tekst? ☐ Is de afbeelding juist gelabeld (itemlabel)?*
Vraag	☐ Is de vraag eenduidig te interpreteren (slechts één antwoord mogelijk)? ☐ Bevat de vraag geen overbodige informatie? ☐ Is de vraag kort, helder, eenvoudig en taalkundig juist geformuleerd (woordgebruik, zinsbouw)?
Alternatieven, sleutel & scoringsaspect (MC)	☐ Sluiten de antwoordalternatieven inhoudelijk aan op de vraag? ☐ Is er slechts één antwoord goed (zijn de afleiders echt fout)? ☐ Zijn de afleiders realistisch (aansluitend bij mogelijke fouten in de berekening)? ☐ Staan de antwoordalternatieven oplopend of op alfabetische volgorde? ☐ Is de sleutel juist en volledig?

	☐ Is er voldoende variatie in sleutels (A, B, C en D) ten opzichte van de andere items? ☐ Staat het juiste scoringsaspect achter de sleutel (niveau- en combinatietoetsen: Domain, Domaintoetsen: Domainonderwerp)?
Invoerveld, sleutel & scoringsaspect (open)	☐ Is de tekst (of geen tekst) voor (bijv. €) of na (bijv. cm) het invoerveld passend? ☐ Is de sleutel juist en volledig? ☐ Is de instelling numeriek van toepassing en zo ja, is/zijn de juiste sleutel(s) ingevuld (bij meerdere sleutels met onderscheiden met|)?** ☐ Staat het juiste scoringsaspect achter de sleutel (niveau- en combinatietoetsen: Domain, Domaintoetsen: Domainonderwerp)?

** zie document ‘Labeling items en afbeeldingen rekenen TOA’*

*** zie document ‘Instructies numeriek’*

Algemene afspraken constructie & screening TOA rekenen	
Weergave constructiedocumenten	
Weergave documenten	▪ Documenten opslaan met datum en eigen initialen achter bestandsnaam. ▪ Titel document toevoegen (bijv. 1F-GE). ▪ Ieder itemformat op een nieuwe pagina. ▪ Itemformats volledig invullen. ▪ Paginanummers toevoegen. ▪ Assetformat volledig invullen. ▪ Assetformat na bijbehorend item toevoegen (op nieuwe pagina).
Aanpassingen n.a.v. feedback	▪ Bij elke aanpassing document opnieuw opslaan, datum wijzigen en initialen toevoegen aan bestandsnaam. ▪ Feedback wordt gegeven d.m.v. opmerkingen en wijzigingen bijhouden. ▪ Bij verwerking feedback wijzigingen verwerken of nieuw voorstel doen d.m.v. opmerking in de kantlijn. ▪ Bij verwerking feedback opmerkingen laten staan zodat de screener kan controleren of de feedback is verwerkt.
Weergave items	
Vorm (kaal/context)	▪ Kale som: som met onbenoemde of benoemde (met maten/eenheden als €, liter etc.) getallen. ▪ Contextsom: items met tekst en/of afbeelding.
Rekenmachine	▪ Rekenmachine is niet toegestaan voor opgaven 0F en 1F. ▪ Rekenmachine is in principe toegestaan bij contextopgaven 2F en 3F, rekenmachine echter alleen aanbieden in TOA als nodig/zinvol om som op te lossen (dus niet bij het aflezen van een grafiek, het benoemen van een meetkundig figuur etc.). ▪ Rekenmachine in principe niet toegestaan bij kale opgaven 2F en 3F, tenzij de berekening zelf niet het doel is (bijv.: items onder volgorde bewerkingen).
Tekst	▪ We nemen geen introductie op in de vorm van ‘Lees de tekst.’, dit spreekt voor zich en is dus overbodige ballast. ▪ Context zo functioneel mogelijk, maar altijd herkenbare situatie die aansluit bij de belevingswereld van de leerling. ▪ Tekst voor zover mogelijk aanbieden in gewone tekst (liever dan bijv. in opgemaakte afbeelding van krantenbericht), zodat de voorleesfunctie toegepast kan worden.

	▪ Tekst voor zover mogelijk geheel boven afbeelding aanbieden i.v.m. met nieuwe weergave in twee kolommen in TEAS. ▪ Tekst altijd in de 3e persoon, niet in 'je' vorm (tenzij in betekenis van men). ▪ Zinnen zoveel mogelijk in de vorm: onderwerp, persoonsvorm, overig. ▪ Tekst achter elkaar doorschrijven, nieuwe alinea met witregel (dus geen tekst op nieuwe regel, tenzij er sprake is van een opsomming). ▪ Opsommingen: woorden/zinsdelen met kleine letter en puntkomma, zinnen met hoofdletter en punt. ▪ Geen bestaande merknamen gebruiken, tenzij niet anders mogelijk.
Vraag	▪ Vraag altijd in vraagvorm. ▪ Vraag bij voorkeur kort introduceren (niet als inleidende tekst overbodig wordt, dit inhoudelijk niet mogelijk is zonder richting te geven aan het antwoord en bij kale sommen). ▪ Bij kale getalsmatige som spatie tussen verschillende onderdelen: 1 + 1 =. ▪ Kale omrekensom in de vorm: Hoeveel procent is ½ deel? (niet meer ½ deel = …%). ▪ Bij vragen met eenheden de gevraagde eenheid benoemen in de vraag (Hoeveel procent is ½ deel?, Hoeveel euro krijgt hij terug?). Voor omtrek, oppervlakte en inhoud hoeft de eenheid alleen benoemd te worden als er sprake is van een omrekening (anders: Wat is de oppervlakte van het veld?). ▪ Onderdelen in vraag die benadrukt moeten worden <u>onderstrepen</u>, bijv. als er iets anders gevraagd wordt dan in de context aan bod is gekomen of als het risico's bestaat dat de leerling ergens overheen leest ('Hoeveel <u>procent</u> is dat?', 'Hoeveel moeten ze <u>samen</u> betalen?').
Algemene afspraken constructie & screening TOA rekenen (vervolg)	
Weergave items (vervolg)	
Figuren	▪ Voor ieder figuur vul je een assetformat in, dat je toevoegt aan het document op de pagina na het bijbehorende itemformat. ▪ Geef elke afbeelding het itemlabel mee, zonder de punten (Wv15-524_301, indien meerdere afbeeldingen in een item, dan _1, _2 etc. aan assetlabel toevoegen). ▪ Afbeelding moeten functioneel zijn (niet enkel illustratief), dat wil zeggen dat de afbeelding nodig is om de vraag te beantwoorden of een deel van de context vervangt). ▪ We nemen geen introductie op in de vorm van 'Bekijk de afbeelding.', dit spreekt voor zich en is dus overbodige ballast. ▪ Afbeelding inhoudelijk inleiden of introduceren in je-vorm: 'Je ziet …', 'Je ziet hieronder zie je…' of 'Je ziet in de grafiek/tabel...'. ▪ Geen bestaande merknamen gebruiken, tenzij niet anders mogelijk. ▪ Voor afspraken figuren (waaronder technisch tekenwerk als tabellen, diagrammen, plattegronden en kaarten), zie document 'Checklist screening assets'.
Woordformule	▪ Woordformule in de vorm van startgetal en groeigetal: prijs per persoon = € 12,50 + € 2,50 x aantal uur (niet prijs per persoon = € 12,50 + € 2,50 per dag).
Getallen, maten en eenheden	▪ Getallen uitschrijven in tekst als het om algemene aantallen t/m 10 à 20 gaat ('twee vriendinnen', 'eens in de veertien dagen'), getallen opnemen in tekst als het gaat om een hoeveelheid of maat ('8 keer', '10 cm', '2 kg'). Richtlijn: getallen t/m 12 uitschrijven in tekst. ▪ Getallen t/m 20 en tientallen ook uitschrijven in sleutel (12|twaalf).

	▪ Getallen vanaf 10.000 in tekst/figuur met punt tussen duizendtal en honderdtal, onder 10.000 aaneengesloten (8000). ▪ Notatie decimale getallen met komma, niet met punt (2,5). ▪ Notatie breuken met verticale streep voor de invoer in de TOA, dus $\frac{1}{2}$ noteren we als |1|2|, 1 $\frac{1}{2}$ noteren we als 1 |1|2|. ▪ Altijd kleinste weergave breuk gebruiken ($\frac{1}{2}$ i.p.v. $\frac{2}{4}$). ▪ Spatie tussen getal en eenheid (€ 2,25, 5 km). ▪ Bij eenheden in de vorm van een maat (met een getal ervoor) gebruiken we de afkorting (2 m, 10 l, 500 g). Achter het invoerveld staat dus altijd de afkorting. Als het geen maat is (er dus geen getal voorstaat) dan schrijven we deze uit (‘Hoeveel meter is er nodig?’). Dit geldt alleen voor de officiële afkortingen, dus niet voor tijden. ▪ Voor de eenheden van tijden (uur, minuten en seconden) gelden geen officiële afkortingen, dus deze schrijven we altijd voluit, ook als er een getal voor staat. ▪ Afkortingen dagen: ma, di, wo, do en vr. ▪ Afkortingen snelheden: km/u en m/s. ▪ Notatie tijden: ##:## uur. ▪ Notatie gehele bedragen: € ##,-.

Weergave vraagtypes	
Meerkeuze	▪ Antwoordalternatieven aanbieden op alfabetische volgorde. ▪ Indien breuken in antwoordalternatieven, dan op volgorde van laag naar hoog op basis van teller-noemer. ▪ Antwoordalternatieven: zinnen met hoofdletter en punt, woorden/zinsdelen met kleine letter en zonder punt.
Open	▪ Afrondinstructie: '*Rond je antwoord af op [aantal uitgeschreven] decimaal/decimalen*' (bijv.: 'Rond je antwoord af op één decimaal.' of 'Rond je antwoord af op een geheel getal.'). ▪ Alleen maten en eenheden voor (€) of na (m, l, kg, etc.) invoerveld; geen 'bakjes', 'keer' etc.), afkortingen gebruiken. ▪ Instellingen numeriek: zie document numeriek.

Afwijkende constructieafspraken IEP Eindtoets rekenen	
Referentieniveaus	
Samengesteld e grootheden	▪ Samengestelde grootheden vallen bij 2F onder het Domain Verhoudingen (7. Samengestelde grootheden), maar bij 1S onder het Domain Meten en Meetkunde (1. Maten).
Weergave rekenopgave	
Vorm (kaal/context)	▪ Kale som: som met onbenoemde getallen (maten/eenheden zijn niet toegestaan). ▪ Contextsom: som met benoemde getallen (maten/eenheden als €, liter etc.), items met tekst en/of afbeelding.
Rekenmachine	▪ Geen rekenmachine toegestaan bij 0F, 1F, 1S en 2F.
Sleutel	▪ Sleutels onderscheiden met spatie|spatie en in alle mogelijke varianten, inclusief de eenheid die al voor of na het invoerveld staat. Bijv.: 5 | vijf | 5 m | vijf m | 5 meter | vijf meter 2,50 | 2.50 | € 2,50 | € 2.50 | 2,50 € | 2.50 € | 2,50 euro | 2.50 euro
Schrijfwijze getallen & eenheden	
Getallen, maten en eenheden	▪ Notatie breuken als $\frac{1}{2}$, met behulp van *Invoegen – Vergelijking – Breuk.* [4]

[4] Alleen bij opgaven in gedrukte toetsboekjes, niet als de opgaven worden ingevoerd in de TOA.

Appendix B: Interviews pilot numeracy tests 2F

Doel van de interviews
Optimalisatie van de samengestelde rekentoets voor leerlingen van vmbo en studenten van het mbo niveau 3.

Thema's

1. Studentkenmerken waar rekening mee moet worden gehouden
2. Voorkeur voor een variant (tekst/beeld/tekst & beeld)
3. Inhoud van de toets

Introductie
Korte introductie over het doel van het interview. Praktische zaken zoals het opnemen van het interview en de anonieme verwerking van de gegevens worden besproken.

Thema 1: Studentkenmerken waar rekening mee moet worden gehouden

1. In welk leerjaar zit je?
2. Welk niveau doe je?
3. Wat is je profielrichting of welke opleiding volg je?
4. Wat is je cijfer voor rekenen?
5. Heb je dyslexie?
6. Heb je dyscalculie?
7. Wat is je leeftijd?

Thema 2: Voorkeur voor een variant (tekst/beeld/tekst & beeld)

8. Heb je een voorkeur voor één van de drie varianten in de toets (tekst/beeld/allebei)?
   Waarom wel of niet?
9. Is dit voor alle opgaven hetzelfde?
   Waarom wel of niet?

Thema 3: Inhoud van de toets

10. Wat vind je van de instructies voorafgaand aan de toets?
11. Wat vind je van de instructies in de toets?
12. Hoe vind je de gebruiksvriendelijkheid van de toets? Denk bijvoorbeeld aan het gebruik van de rekenmachine, de grootte van de plaatjes, de leesbaarheid van de tekst en de getallen.
13. Zitten er onduidelijkheden of moeilijke woorden in de toets?
14. Wat vind je van de duur van de toets?
15. Heb je suggesties voor verbetering?

Bedanken
Bedankt voor het meewerken aan dit onderzoek.

Appendix C: Informed consent form

Beste deelnemer,

Door mee te werken aan dit onderzoek, help je toetsing in Nederland te verbeteren en kun je oefenen met opgaven op niveau 2F. Dit is een onderzoek in opdracht van de Universiteit Twente en Bureau ICE. Het onderzoek bestaat uit een rekentoets met een korte vragenlijst.

De gegevens en resultaten van het onderzoek zullen alleen anoniem en vertrouwelijk aan derden bekend gemaakt worden. Voor vragen kun je mailen naar onderstaand mailadres.

Bedankt voor je hulp en veel plezier met de toets!

Met vriendelijke groet,

Lynn Buschers
Masterstudent Educational Science and Technology
l.j.buschers@student.utwente.nl

Toestemmingsverklaringformulier (informed consent)
Titel onderzoek: The influence of the presentation of contextual numeracy problems on student performance in (pre)vocational education
Verantwoordelijke onderzoeker: Lynn Buschers

In te vullen door de deelnemer
Ik verklaar op een voor mij duidelijke wijze te zijn ingelicht over de aard, methode, doel en [indien aanwezig] de risico's en belasting van het onderzoek. Ik weet dat de gegevens en resultaten van het onderzoek alleen anoniem en vertrouwelijk aan derden bekend gemaakt zullen worden. Ik begrijp dat audiomateriaal of bewerking daarvan uitsluitend voor zal worden gebruikt. Mijn vragen zijn naar tevredenheid beantwoord.

Ik stem geheel vrijwillig in met deelname aan dit onderzoek. Ik behoud me daarbij het recht voor om op elk moment zonder opgaaf van redenen mijn deelname aan dit onderzoek te beëindigen.

Naam deelnemer:
…………………………………………………………………………………....………………..

Datum: ………..…… Handtekening deelnemer: …...…………………………………………………….

In te vullen door de uitvoerende onderzoeker
Ik heb een mondelinge en schriftelijke toelichting gegeven op het onderzoek. Ik zal resterende vragen over het onderzoek naar vermogen beantwoorden. De deelnemer zal van een eventuele voortijdige beëindiging van deelname aan dit onderzoek geen nadelige gevolgen ondervinden.

Naam onderzoeker: Lynn Buschers
Datum: …………..…… Handtekening onderzoeker: ...…………………………………………………..