Relations between reading comprehension and mathematical word problems of varying complexities

A cross-sectional study of the Norwegian national tests aiming to identify relations between reading comprehension and word problems, and if task properties affect student sub-groups differently.

JONATHAN HELFJORD

HANS KRISTIAN
Hans Kristian Nilsen

KIRSTEN
Kirsten Bjørkestøl

University of Agder, 2021
Faculty of Engineering and Science
Department of Mathematical Sciences

Abstract

This thesis aims to investigate the relationships between reading comprehension and mathematical word problems of varying complexity. There exist numerous studies on this relationship (Abedi & Lord, 2001; Bergqvist et al., 2018; Cummins et al., 1988; Greer, 1997; Maagerø & Skjelbred, 2010; Martiniello, 2008; G. Nortvedt, 2009; Verschaffel et al., 2000; Vilenius-Tuohimaa et al., 2008), with differences in both thematic and research design levels. Most use a quantitative approach to their research question, while some complements with students' perspectives, thoughts, and solution processes.

My study is a quantitative, cross-sectional study examining reading comprehension and word problems through the context of the Norwegian national tests, with 5854 anonymous participants from eighth grade, all across Norway. I approach my research question through multiple analyses; a correlation analysis between reading comprehension and numeracy tasks, and logistic regression analyses for selected numeracy tasks with reading comprehension as a predictor. In the correlation analysis, I also investigate the sub-question of *whether task properties affect student sub-groups differently, based on their reading comprehension level.* This was done be categorizing the participants, and subsequently performing new correlation analyses between reading comprehension and all numeracy tasks for each sub-group. The results are then discussed in the light of the framework of cognitive demands of mathematical tasks (Stein & Smith, 1998).

There are three main findings highlighted in my results. Firstly, I identified a tendency for students with low reading comprehension level to benefit from said ability on different tasks than the others, which was not unanimously tied to the complexity level, but rather a combination including the cognitive demand and the difficulty thresholds of each task. Secondly, my data supports that the top performers in reading comprehension attain stronger relationships with word problems when the complexity level and the cognitive demand of a task increase. Thirdly, the correlation analysis revealed that some tasks possessed a relatively strong relationship with reading comprehension. The logistic regression models showed that using reading comprehension as a predictor increased the accuracy of these models better than for the remaining tasks. Also, the tasks had both high discrimination and difficulty thresholds, showing that categorizing word problems based on the mathematical complexity is insufficient to fully capture the nuances of each task.

Table of contents

1. Introduction

1.1. Background and motivation

Throughout my teaching education I have met students complaining about word problems, stating numerous variants of "I do not understand what they are asking". This area of difficulty for students has always intrigued me, so when it was time to write my master thesis, I knew that this is a subject I wanted to investigate. I have always been fond of statistical analysis, so I desired to combine the mathematical word problems with the quantitative nature of statistical analysis. To explore this area, I found that the magnitude of the national tests fit my purpose.

Word problems is a challenging part of the mathematical curriculum, with students scoring significantly lower than on basic arithmetic problems (Cummins et al., 1988). They are also widely used in textbooks and assessments, and students are faced with word problems throughout their mathematical journey. Still, even though students are continually exposed to these problems, reports suggest that students experience them as complex, difficult to understand and lacking of real world connections (Greer, 1997). There has been conducted extensive research to answer why students experience such difficulty, pointing at a variety of factors. Because of the many underlying factors, there are several branches of research on word problems and reading comprehension. Still, most researchers agree that there exists a relationship, and that is what I want to elaborate on. I found a gap in the research that discuss the varying complexity levels of word problems, and it seemed to be limited with studies on sub-groups of students based on reading comprehension levels. I wanted to address this, and that motivated the research question "*Which relations exist between reading comprehension and mathematical word problems of varying complexities?"*, along with the sub-question "*how do task properties affect subgroups of students, based on their reading comprehension level, differently?"*.

My thesis will expand on the existing literature of the relationship between reading comprehension and the solving of mathematical word problems. This will be done by separating the word problems given on the national tests to eight graders on complexities and examine differences in those new groups. Moreover, I will split the student group into different levels of reading comprehension and investigate how the different level groups perform on the complexities of the mathematical word problems. Although the analysis is based around the context of the national tests, I will present findings that can have implications for how we perceive word problems, and how they affect student sub-groups differently.

2. Theoretical background

My theoretical background serves multiple purposes: Firstly, it provides insight in some key mathematical concepts and why they are relevant to my thesis, namely theories on how students acquire knowledge and aspects of the basic skills measured in the national tests. Secondly, it elucidates differences in mathematical tasks, and helps me distinguish the features of word problems. Thirdly, it provides a framework for analyzing the properties of the tasks through Stein & Smith (1998). All three purposes will subsequently help me investigating my research question and position myself into the field of mathematics education.

2.1. Acquiring mathematical knowledge

By acquiring mathematical knowledge, I mainly refer to Hiebert's (1986) procedural and conceptual knowledge. One of the most prominent studies on mathematical knowledge is the five strands by Kilpatrick et al. They consider *mathematical proficiency* a necessary skill to learn mathematics successfully, and they divide this ability into five components: *conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive disposition* (Kilpatrick et al., 2001). The authors believe that neither component is sufficient, when isolated, for learning mathematics, hence the intertwined strands. Make note that the components include elements from both procedural (procedural fluency and productive disposition) and conceptual (conceptual understanding and adaptive reasoning) knowledge. Thus, Kilpatrick et al. (2001) does not think that either should be the dominant component of acquiring knowledge.

Other researchers disagree about whether procedural knowledge is necessary for acquiring conceptual knowledge, or vice versa. Rittle-Johnson (2017) have previously proposed an iterative process, where both procedural and conceptual knowledge influence each other, and she argues that they are intertwined and cannot be interpreted as a unidirectional relationship (Rittle-Johnson, 2017). Moreover, some found that there might be individual differences in how students draw from both procedural and conceptual knowledge (Hallett et al., 2010). They state their position closely to the previous research done by Rittle-Johnson, that students might learn through procedural-first or conceptual-first. However, Hallett et al. (2010) makes a distinction in that they do not assume that one type of knowledge inherently leads to an increase in the other. Moreover, on the discourse regarding procedural or conceptual knowledge being the most influential for acquiring mathematical knowledge, there are arguments for some middle ground where neither should be relinquished. Neither fully capture the full extent of mathematical knowledge, they fill different roles in a teaching discourse and combining them would lead to a holistic approach (Sfard, 1998). Thus, even though new literature elucidates the strength of conceptual knowledge, it should be important not to discard procedural knowledge completely.

It is important to remember that my research does not give indications of students solution strategies, thought processes or their previous knowledge of the subjects assessed on the national tests. Also, the degree of how much students have previously engaged in conceptual tasks will vary greatly, depending on the teaching methods of the teachers. However, the concepts of procedural and conceptual knowledge are closely tied to concepts I use in assessing tasks later in my analysis. Hence the importance of elucidating this as a theorical background, even though I cannot directly assess it through my dataset.

2.2. Basic skills

My research question seeks to investigate reading comprehension and the solving of word problems, so in order to do that I first wanted to clarify how I considered reading comprehension and numeracy, and how it compares to the interpretation by The Directorate of Education, henceforth called UDIR. Basic skills are defined in the Norwegian national curriculum as “essential tools for development and understanding. It is also a necessity for showing knowledge."(Utdanningsdirektoratet, n.d.-b). The basic skills include reading comprehension, numeracy, and written, oral and technical ability and are thought to be taught in every subject. Moreover, UDIR states that some subjects have wider responsibilities for teaching the basic skills, e.g., mathematics having majority of the responsibility of providing the students with development in numeracy.

To further specify the basic skills, UDIR has defined a set of properties for each skill and what defines the level of ability for each property. Because of the focus of my research, I chose to omit the written, oral, and technical abilities and focus on reading comprehension and numeracy. I recognize that the translated names reading comprehension and numeracy have wide connotations and interpretations across multiple fields of study, so as a reader, keep in mind that my interpretation coincide with the interpretations of UDIR. Furthermore, they present the descriptions of each aspect of reading comprehension and numeracy, along with five levels of each aspect clarifying what defines each level. To get a sense of the structure I included the description of five levels of the *interpret* aspect of reading comprehension:

Table 1 - Levels of the aspect interpret.

Level of the aspect interpret	**Description**
Level 1	Use your own words and draw simple conclusions from information in texts.
Level 2	Identify core ideas and understand connections which are explicitly defined in texts.
Level 3	Understand implicitly written information in texts.
Level 4	Understand ambiguity, identify contradicting information and information opposed to the expected.
Level 5	Show a detailed and comprehensive understanding of complex texts, and can structure and draw conclusions based on implicit information.

This is done for the four aspects of reading comprehension: *preparation, finding information, interpreting, and reflecting*. Moreover, the same structure applies to the four aspects of numeracy: *recognize and describe*, meaning that students should be able to recognize situations where numeracy can be applied, *applying strategies*, that students are able to apply their knowledge to mathematical problems; use appropriate tools and strategies, *reflect and consider*, that students can reflect and validate their solutions, *and communicate,* the last one not being assessed in the national tests (Utdanningsdirektoratet, 2017). However, it is important to remember that the basic skills are not constrained to one subject – numeracy can be taught through discussions of graphs and tables in social studies and linguistic subjects, and literacy can be taught through word problems in mathematics. This creates a base for my research, but before discussing word problems I wanted to elaborate on mathematical tasks and their properties.

2.3. Different kinds of mathematical tasks

The preliminary investigations into the national tests revealed that I needed a framework for distinguishing each numeracy task based on their properties. One way was through the framework by Stein & Smith (1998), which categorizes mathematical tasks as having levels of cognitive demand, with some innate properties of each task types. From *memorization tasks*, *procedures without connections*, *procedures with connections,* to *mathematising* tasks, the cognitive effort requires increases with the categorized levels. Albeit being exclusive categories with specific properties tied to each, the authors emphasize that it should not be considered a rigid structure, but rather a framework for reflection (Stein & Smith, 1998). This framework helped me categorizing the numeracy tasks, which laid the foundation for some interesting findings, which will be discussed in chapter *6* and *7.*

This way of distinguishing tasks into categories is closely related to the procedural and conceptual knowledge discussed in chapter *2.1*, with tasks having lesser cognitive demands typically testing procedural knowledge. It is shown through studies that students engaging in such tasks tended to have insufficient knowledge when faced with unfamiliar tasks (Boaler, 1998). Her study highlighted differences in procedural and conceptual approaches to mathematics in two schools, and the school engaging in open tasks with high cognitive demand developed a conceptual understanding that enabled them to apply strategies in unfamiliar situations. Moreover, they indicated a more positive view about mathematics (Boaler, 1998). This supports the use of authentic, open tasks in the mathematics classroom compared to the narrow, procedure-focused ones, for a deeper understanding of the mathematical topics.

The open, rich tasks have properties that are worth examining. Firstly, the property *open* is stated in the phrasing. An open task requires students to make choices about their solution strategy, which mathematical knowledge to apply, and that they can formulate arguments as to whether the answer seems reasonable. This shows that tasks can be open in the start (freedom of solution strategy) and in the end (no definitive answer, opens for discussions of what is considered correct and why). These requirements coincide with the properties of numeracy as a basic skill, and provides support for the increased focus on a conceptual approach to understanding in the Norwegian national curriculum.

In addition to the cognitive demand framework, I needed to justify that the tasks could be considered word problems, therefore being a valid dataset for my research question. The distinction of word problems is addressed by Blum & Niss (1991), who wrote that problem solving can be approached through a pure mathematical context, or through some real-world scenario but still within a mathematical context, hence “dressed up”. Moreover, they describe the modelling problems as an applied problem solving process where the solver needs to identify a real problem situation, then simplify, structure and justify a mathematical model that could fit (Blum & Niss, 1991). Newer research speaks to the same distinction between tasks, but with a slight renaming to intra-mathematical tasks, word problems and modelling problems (Schukajlow et al., 2012).

It is easy to assume that intra-mathematical tasks, without any real-world connections, should be considered tasks with low cognitive demand. Likewise, that modelling tasks are tasks with high cognitive demand. However, this is a simplification that insufficiently categorize mathematical tasks. For example, a task asking to investigate patterns in frequency of prime numbers is purely mathematical, but could have a vast number of solution processes, so constraining it to one correct procedure would not be feasible. To connect this example with the framework of Stein & Smith (1998), the task asks students to assess a strategy, apply appropriate mathematical knowledge and draw conclusions based on their findings. These requirements situate the task in the threshold between *procedures with connections* and *mathematising*, showing that intra-mathematical tasks do not necessarily mean low cognitive demand. Thus, the separation between intra-mathematical tasks, word problems and modelling problems should not be constricted to low, medium, and high cognitive demand, respectively, but rather as an independent structure to the cognitive demands.

To conclude, I found that the framework of cognitive demand for tasks by Stein & Smith (1998) would be beneficial in helping me distinguish between the different numeracy tasks, based on their properties. It provides a foundation for the discussion of my results, and adds a layer to distinguish tasks that are considered having same properties by UDIR. Moreover, I presented a way of considering mathematical tasks through their connection to the real world. In this lies the distinction between word problems and other mathematical tasks, which is essential for my argument of why the national tests are considered valid for investigating word problems. To further strengthen this argument, the next sub-chapter presents defining properties of word problems.

2.4. Word problems

In order to justify why the numeracy tasks are a valid way of examining word problems, I needed to clearly define what a word problem is. Maagerø and Skjelbred (2010) states that mathematical word problems have several distinct properties separating them from texts in other subjects; they have a high frequency of multimodality, are information dense and often include specific terminology required by the reader to be familiar with (Maagerø & Skjelbred, 2010). This speaks to word problems being complex structures with a high degree of cognitive demand, coinciding with *mathematising* tasks (Stein & Smith, 1998) as discussed above. Moreover, Verschaffel, Greer and de Corte writes that "a characteristic feature of word problems is the use of words to describe a (usually hypothetical) situation." (Verschaffel et al., 2000). They also make a distinction from simple algebraic calculations formulated into words, saying that tasks such as "what do you get if you subtract 3 from 8" does not fit the criteria of a word problem. Thus, it could seem like word problems, as defined by Verschaffel et al. (2000) possess some similar properties to rich tasks which was introduced in the previous sub-chapter. However, word problems can have different levels of cognitive demand and can be considered memorization tasks or procedures without connections, not only mathematising tasks. Therefore, it cannot be stated whether word problems share a common feature of having low or high cognitive demand, as the variations in tasks can differ greatly.

Based on the above, I assume that word problems are not considered intra-mathematical, the concept defined previously by (Schukajlow et al., 2012). Even though intra-mathematical tasks can be cognitively demanding, asking students to explore complex mathematical structures, they still do not possess the property of being connected to a real-world scenario. Continuing on this, the distinction between word problems and modelling problems should also be accounted for. An influential model used to describe modelling problems is by Blum & Weiß (2007), as cited in (Blum & Ferri, 2009). It describes how students assess a real world problem, convert it to mathematical terms and make a simplified model, test the results and then validate how the model fits the reality (Blum & Ferri, 2009). The most glaring difference between word problems and modelling problems is how they approach the mathematical content; word problems are considered mathematical of nature, described in a real-world context. Therefore, the mathematical answer is the ultimate goal of the task. For modelling problems however, mathematics is used to simplify an existing situation, to help solve a problem. Thus, the goal of a modelling task is not the mathematical answer, but rather the considerations of how well the model fits compared to reality.

This sub-chapter provided a base for separating word problems from intra-mathematical tasks and modelling problems, along with some general properties of word problems. It is clear that every numeracy task in the national tests is situated around a real-world context, excluding any of them from being intra-mathematical tasks. Moreover, none of them possess the complex structures of modelling problems. Consequently, I consider every numeracy task to be a word problem and will treat them as such through the analyses. With this I conclude the elaboration of word problems in the didactical research of mathematics, continuing to the other skill I am assessing through the national tests, namely reading comprehension.

2.5. Reading comprehension

There is a plethora of research on reading, and more specifically reading comprehension. This is captured in the quote “there is no theory of reading, because reading has too many components for a single theory” (Perfetti & Stafura, 2014). This master thesis, even though being interdisciplinary, are mainly focused on the mathematical aspect of word problems through the tasks in the national tests, and how they relate to reading comprehension. Consequently, the theoretical background also centers around a mathematical perspective on reading. However, to display the differences in how reading comprehension is perceived I wanted to present some general theories of reading comprehension, albeit shallow compared to what could be expected from a pure linguistic study.

With a generalized, broad scope reading comprehension can be interpreted by two different branches. The first one assumes that text is contextual, meaning that the reader’s situation affects their comprehension of the text (Van Dijk & Kintsch, 1983). When examining reading comprehension in this scope, the reader’s experiences and knowledge influence how they understand the text, thus constructing their own understanding. This is further supported by other researchers as schema theory, that text does not carry meaning by itself (Carrell &

Eisterhold, 1983). The other branch emphasize that comprehension can be developed without regard to the discourse context, and a theory of describing this is called the Construction-Integration model (Kintsch, 1988). Opposed to the first branch, this assumes that the reader create knowledge through the linguistic properties, and that these are integrated into the reader's knowledge base, to be able to understand the full text. Thus, the C-I-model assumes that knowledge is transferred from author to reader.

Building on the two concepts, Flood & Lapp (1990) summarized research on how competent readers actively construct meaning through interacting with the words on the page, integrating new information with their pre-existing knowledge. Also, research supports that a reader's prior knowledge, experience, attitude, and perspective determine the way text is understood (Flood & Lapp, 1990). Thus, there are research supporting both branches, and neither can be seen exclusively as a tool to analyze a reader's understanding of a text. They continue by identifying some methods usually applied by competent readers in their reading process, including but not limited to the following: building background by activating appropriate knowledge, sets purposes, monitors comprehension, integrates new concepts and makes applications of the ideas in the text.

As shown, there is a multitude of ways to conceptualize reading comprehension, and these examples work as proof of how the different perspectives can alter the theories in opposite directions. While some theories emphasize the contextual situation for the reader, implying that a text does not contain knowledge until being read, others assume an approach where reading comprehension is a de-coding process of the linguistic properties. The vast differences motivate research into the next sub-chapter, how mathematics view reading comprehension.

2.5.1. Reading comprehension in mathematics

In order to investigate the connection between reading comprehension and word problems, it is insufficient to establish theory on reading comprehension in general. This notion is motivated by the quote from Maagerø & Skjelbred (2010) in *2.4* about word problems in mathematics being different from texts found in other subjects. Therefore, to establish how the mathematical research community views reading comprehension, I will present some general directions of study and how it relates to word problems.

Research revealed that there have been numerous studies on the relationship between reading comprehension and word problems in mathematics (Bergqvist et al., 2018; Cummins et al., 1988; Martiniello, 2008; G. Nortvedt, 2009; G. A. Nortvedt, 2011; Österholm & Bergqvist, 2012; Vilenius-Tuohimaa et al., 2008). Some focused on the linguistic properties, like Abedi & Lord (2001) who discovered that by simplifying the linguistic factor in mathematical word problems, students in low-level math classes, or student with language barriers, performed slightly better (Abedi & Lord, 2001). This is supported by (Martiniello, 2008) who found that students with low English ability scores lower than natural English speakers with equal mathematical proficiency, and that these differences increase parallel to increasing linguistic complexity (Martiniello, 2008). As an argument to why the linguistic properties affect the solution process, Kintsch et al. (1988) observed that minor alterations in wording of mathematical tasks affected the percentage of students correctly solving the task, even though the mathematical concept of the original and altered tasks were the same (Kintsch, 1988). This research, with focus on the linguistic properties, contain some similarities to the C-I model, in that it is about students integrating meaning from the text into their own understanding without taking into consideration the student context. One problematic aspect of reviewing the linguistic properties of mathematical word problems is that analyzing tools often require longer texts (Homan et al., 1994), thus being ineffective for the compact, information-dense word problems.

Other studies examined student strategies when faced with word problems, like (G. Nortvedt, 2009). She used the national tests for a statistical analysis of the correlation between reading comprehension and word problems, and complemented the results with discussions of strategies from a small sample of students. Results revealed that students with proficient ability in reading comprehension tended to accompany proficiency in numeracy, and that the proficient readers were able to adjust their proposed models better than the low-proficiency readers. Similar results were found in another study, where students categorized as good readers (GR) performed better than poor readers (PR) on both mathematical word problems and reading comprehension (Vilenius-Tuohimaa et al., 2008). Even though this study did not discuss strategies like Nortvedt (2009), they both distinguish between levels of reading ability, with a subsequent discussion about differences between the groups.

An interesting aspect of the article by Flood & Lapp (1990) is that the methods applied by the competent reader has many similar features to the description of numeracy as a basic skill, defined by UDIR.

Table 2 - Comparison of competent readers and numeracy

General idea	**Methods by competent readers**	**Numeracy as a basic skill**
Use prior knowledge to identify properties in the text	Build background, set purpose, and check understanding	Recognize and describe
Fitting process between existing and new knowledge	Monitor comprehension and integrate new concepts	Apply their strategies
Control if their understanding seems plausible	Summarize and evaluate, make applications	Reflect and consider

(Flood & Lapp, 1990; Utdanningsdirektoratet, n.d.-b)

This comparison visualizes how the methods applied by competent readers have similarities to the aspects of numeracy as a basic skill. As discussed in *2.2*, one way of measuring basic skills is through the national tests, and the numeracy test aims to determine students' proficiency on these aspects of numeracy. All the 50 tasks in the numeracy test, which I will elaborate on in chapter 3, are considered word problems, based on the theorical background presented in previous sub-chapters. They have specific real-world contexts, and consequently are not considered intra-mathematical tasks. Moreover, they lack the complex structure of a modelling problem. Thus, assuming that the numeracy tasks included in the national tests are considered word problems is considered appropriate, based on the theorical background provided. Combined with the methods of competent readers by Flood & Lapp (1990) the connection between reading comprehension, word problems and the numeracy tasks are established.

2.6. Summary

I started this chapter with presenting some learning theories, namely the *five strands of mathematical proficiency* (Kilpatrick et al., 2001), and a discussion about the views of procedural and conceptual knowledge in mathematical research. The conceptual approach emphasized engaging in mathematical discussions such that students could develop their reasoning, critical thinking, and have a foundation to apply in new situations. These areas coincide with the basic skills that UDIR wants students to develop through their education, which again bear similarities to the strands of mathematical proficiency (Kilpatrick et al., 2001). It is important to remember that I aim to investigate word problems through the national tests, so consequently I did not position myself firmly into one specific learning theory.

They were, however, essential as a foundation for building on theories of mathematical tasks, and how word problems are distinguished from other mathematical tasks. Making that distinction was necessary for determining how I can treat the numeracy tasks in the national tests, and that distinction was made through several frameworks. The cognitive demand of tasks (Stein & Smith, 1998) provided me with a basis for distinguishing between each numeracy task, but in order to justify treating every numeracy task as a word problem I conferred to Blum & Niss (1991) for separating word problems from intra-mathematical tasks and modelling problems. Thus, the theoretical justification of categorizing word problems was established, which lead me to the other skill assessed in the national tests, reading comprehension.

I presented studies on the relationship between reading comprehension and word problems, specifying some of the areas they chose to focus on. Moreover, I identified similarities between strategies applied by good readers and the aspects of numeracy as a basic skill. This showed that, even though being two different disciplines, there are aspects of reading comprehension that enhances numeracy, and vice versa. That notion is supported by basic skills as being thought of as interdisciplinary, not to be taught exclusively in the most approximate discipline (Utdanningsdirektoratet, n.d.-b). Thus, through the theoretical framework I established relevant theories to my research question and identified a framework in which I could evaluate the numeracy tasks in the national tests. I also justified both the choice of the national tests as a measurement tool, and that the connection between word problems and reading comprehension is worth investigating. Before discussing my methodology and strategy for answering the research question, the national tests are presented and discussed.

3. National tests

Following the theoretical background is a description of the structure and aim of the national tests. The structure of the reading comprehension and numeracy tests are elaborated upon, and I discuss how I treated the categorizations of the variables from each test. Ultimately, I present some reflections about why the results from the national tests are considered an appropriate tool for analyzing my research question.

3.1. The purpose of national tests

The national tests aim to:

"Give the schools knowledge about their students' basic skills in reading comprehension, numeracy and English proficiency. The information from the tests is forming a baseline for continuous evaluation and quality control on every level of the school system" (Utdanningsdirektoratet, n.d.-c).

They are conducted in the autumn every year for fifth, eighth and ninth graders, and acts as an analyzing tool on school-, class- and student level to identify development opportunities. It is important to remember that the tests do not fully represent the students' abilities in the basic skills, the results have to be seen in conjunction with other results collected at the school.

3.2. National tests in reading comprehension and numeracy

Before discussing the structure of the national tests, let me reconvey quickly to chapter *2.2* about basic skills. UDIR has defined four aspects of reading comprehension, where neither can be evaluated exclusively to assess a student's reading comprehension ability. One of the four aspects, preparation, are not measured in the national tests, leaving the aspects *find, interpret,* and *reflect* to be assessed.

For the 2020 version of the reading comprehension test, students were given 7 texts with different themes. Each text had six tasks associated with it (the first text had seven), with all three aspects of reading comprehension represented in each text. Most tasks had multiple choice options, where students are given 1 if they answer correctly and 0 if it is incorrect or not answered. Eight tasks had open questions, requiring students to write their answer. For these tasks, UDIR includes an assessment guide, highlighting the threshold between wrong and acceptable answers.

Numeracy is also a multifaceted basic skill, and *recognize and describe*, *applying strategies*, and *reflect and consider* are measured in the national tests. However, as opposed to the reading comprehension test, the aspects of numeracy are not exclusively tied to individual tasks. The reason for this is that the aspects of numeracy combined make up the problem solving process (Utdanningsdirektoratet, 2017). Instead, the 50 tasks in the numeracy test are given complexity levels, from 1 to 5. To make sure that correct answers on high complexity tasks are valued higher than low complexity ones, the logistic regression models of each task include both task difficulty and potential discrimination, which will be elaborated upon in the next sub-chapter.

3.3. Framework of the dataset

In the dataset provided by UDIR there were a total of 96 variables. One dichotomous variable describing gender with the data-entry J for girls (jenter) and G for boys (gutter). Two variables were called “reading comprehension raw score” and “numeracy raw score” respectively, indicating the total amount of points acquired by each student. The last 93 variables were reading comprehension task 1-43 and numeracy task 1-50. Except reading comprehension task 30 with a maximum score of 2 points, the 92 other task variables were dichotomous with 1 for correct answers and 0 for incorrect.

The national tests build on Item Response Theory to evaluate each task and secure consistency through the years of testing. They describe IRT as a tool to “estimate the tasks difficulty unbiased by the participating students, and you can estimate the students ability level independently of the tasks they answered (Utdanningsdirektoratet, 2018b). IRT can be used with several parameters, but the chosen one for evaluating tasks by UDIR includes two parameters: task discrimination and difficulty. The probability formula is:

$$\boldsymbol{P\ (\theta) = \frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}}}$$

Where P is the probability of correctly answering the task, a is the task discrimination, b is the task difficulty, and the constant D is used to approximate a cumulative, normal distribution of the ability (Utdanningsdirektoratet, 2018b). Note that UDIR missed the D*a in the numerator, which appears to be a writing error. The D is caused by the first IRT-models, which was calculated differently than the model used now. Thus, for enabling UDIR to evaluate the tests for different years on the same scale, D is included. For a more in-depth description of the validities, the uncertainty related to the tests, and how IRT enables evaluating changes in test versions for different years, see (Utdanningsdirektoratet, 2018b).

When choosing an appropriate analysis strategy, I had to keep in mind that the raw scores of reading comprehension and numeracy did not fully reflect the students' ability on said basic skills. This was especially relevant for the logistic regression analysis, which bears some resemblance to the models produced by IRT.

3.4. How my research relates to the national tests

So far, I have discussed learning theories in mathematics, different kind of mathematical tasks and the position of word problems. If I examine the numeracy tasks in the national tests with the scope of Verschaffel et al. (2000), all the tasks are considered word problems due to them having a specific real-world context, elevated above a pure mathematical concept. The reading comprehension test is a measure of the students reading comprehension ability, so by using the national tests in my research I gained several benefits: coherency between the reading comprehension and numeracy test, possibility of a large dataset with a random sample collection, and the structure enabled me to assess subgroups' performance on complexity and task level. Therefore, the national tests in reading comprehension and numeracy are considered appropriate for investigating my research question.

4. Methodology

My methods chapter includes consideration about my chosen research paradigm, design, and the collection of data. Due to my thesis having two methods of analysis; correlation and logistic regression, I chose to have the methodology specific for the analyses included in their own chapters. That made for transparency in the model considerations, validity and reliability of the analyses and ease of backtracking in case readers wanted to confer with the methodology while studying results.

4.1. Research paradigm

To recall, I wanted to investigate *the relations existing between reading comprehension and mathematical word problems of varying complexity,* and subsequently *if task properties affected student subgroups, based on their reading comprehension level, differently*. Because of my desire of findings beyond the scope of the national tests, a qualitative approach could limit my possibility of generalizing the results (Bryman, 2012, p. 390). It would also be challenging to engage in the properties of the tasks without discussing the opinions of the students I would have interviewed, thus shifting the core focus towards their experience.

There were some compelling arguments for a quantitative approach. According to some, it falls into a positivist research perspective (Cohen et al., 2017), depending on quantitative data and assuming that there exist objective truths to be discovered. An epistemological standpoint of positivism is further described in Bryman (2016) as a “position that advocates the application of the methods of the natural sciences to the study of social reality and beyond” (Bryman, 2012, p.28). If I were to exclusively rely on the data collected and interpreted through the analyses, my thesis could have been considered positivistic. However, there are some critiques to the nature of positivism, i.e., that they assume the researchers conceptualization of reality to reflect that reality (Bryman, 2012,p. 29). I think it would be hubristic to assume that my conceptualizations fit reality perfectly, so consequently I position myself in the critical realism epistemology, acknowledging the simplifications in my study compared to reality. When discussing implications and reasons, I move slightly away from the objective truth of the statistical results.

A quantitative approach assumes that it is possible to collect data and interpret the results to gain new understanding about the underlying constructs. Also, the results originating from a quantitative study could have implications extending past the actual test subjects, depending on the sample. Moreover, a quantitative approach normally includes the assertion of hypotheses and the subsequent testing of them, following a set of stages through strict guidelines (Cohen et al., 2017. p.519). Based on these considerations, I found that a quantitative approach would yield quantifiable results regarding the relationship between reading comprehension and numeracy. I wanted to investigate the underlying properties of the task, uninterrupted by the participants engaging in the tasks, so I determined that a quantitative approach would be the better fit.

4.2. Research design and collection of data

As mentioned in the introduction, word problems have always been a fascinating subject for me. Through my internships at various schools, I found that many students struggled working with them, and I asked myself why that was. After some research into the word problems domain, I realized that reading comprehension was found to be closely related to word problems. My educational background in linguistic (Norwegian) and mathematical subjects supported the decision of investigating this interdisciplinary question further. After reading about the strategies employed by researchers in previous studies, and what they aimed to investigate, I noticed that it lacked depth in research on differentiating word problems. Based on this, I started the process of formulating a research question. This strategy of identifying potential in the theory and subsequently develop a research question coincides with how quantitative studies often emerge, according to Bryman (Bryman, 2012, p.161).

With a general idea and a research question, the next step was to consider the research design. I considered a cross-sectional design to be preferable, because I wanted to investigate a large number of cases and establish relationships based on my findings. Moreover, Bryman (2012) wrote that replicability and the external validity is strong in a cross-sectional design. It is easy to replicate the research scenario because the number of participants, instruments used, and the implementation of the data collection can be reviewed in the methods section. (Bryman, 2012). Moreover, a cross-sectional study can identify population-wide features, and if the sample is sufficiently large, can enable inferential statistics to be applied. (Cohen et al., 2017).

This strengthened my decision for this approach, because I wanted to investigate subgroups in the student sample. Thus, I needed to base my analysis on a dataset that incorporates a large number of participants, unbiased sample collection, and with a transparent process. An ideal candidate for this was the national tests, so I initiated contact with UDIR to ask permission for a dataset. Another benefit of using the national tests, is that they help strengthen the internal validity of my results through the rigorous sampling process, standardized tests, and that almost all students are required to participate.

After initiating contact, I presented my research question and explained my desires for the dataset. I asked for a large number of participants (> 1000), random sample collection, responses from the eighth grade, and that the scores for each task was specified for both the reading comprehension and the numeracy tests. Also, I had to make sure that even though the student responses were anonymous, I could connect each student’s scores on reading comprehension and numeracy. They generously agreed to apply me with the dataset, and I received a sample of the 2020 dataset. To make sure we agreed the format, they sent a smaller dataset beforehand, in case I wanted alterations made to the final dataset.

The final dataset included the responses of 5854 participants in the eighth grade, sampled randomly from all over Norway, which is slightly less than 10% of the total student mass in eighth grade (Statistisk sentralbyrå, 2020). They specified that there were no clusters of responses, such as including all responses from specific counties. The dataset contained 96 variables – gender, raw scores on reading comprehension and numeracy, the 43 tasks in reading comprehension and the 50 tasks in numeracy.

4.3. Analysis strategy

To assess my research questions, I referred to the relevant studies found in the theoretical framework. Studies with similarities, like (G. Nortvedt, 2009; Vilenius-Tuohimaa et al., 2008) used a variety of statistical approaches to their research, i.e., correlation analysis and confirmatory factor analysis. Because of my limited timeframe, and previous familiarity with correlation analyses, I chose that as a method of assessing my research question. Moreover, regression models would yield me results similar to those produced by the IRT analysis by UDIR, and they would help me in examining task properties. Thus, I found a correlation analysis, complemented by logistic regression models for specific tasks, analyzed with SPSS, to be an accessible and valid way of researching my topic.

I was unsure about which correlation model would best fit the dataset. When testing for correlation the Pearson correlation immediately comes to mind, though there are some limitations in the dataset that makes it an improper fit for this dataset. My variables are treated as discreet, while Pearson correlations require variables being measured on an interval or ratio scale, that is being continuous (Schober et al., 2018). Thus, the Pearson correlation were subsequently considered a bad fit for this dataset.

Earlier studies have made use of Spearman Rank correlation to test correlation between reading comprehension and numeracy (G. Nortvedt, 2009; Österholm & Bergqvist, 2012). A difference between Pearson and Spearman is the rank system used by the latter that assigns ranks to each value and perform a correlation analysis on the ranks, hence it can be beneficial to apply where the dataset is not normally distributed. Kendall’s Tau-b correlation test is another test using the ranks, but it is not reported as often as Spearman’s Rho. For my dataset, the coefficients of Spearman’s Rho and Kendall’s Tau-b appeared nearly equal. Thus, in the analysis of the total student group, results from the two correlation tests are indistinguishable. However, to answer the research sub-question the student group will be separated into smaller entities. Therefore, I considered Kendall’s Tau-b to be better suited than Spearman’s Rho, due to the former having smaller gross error sensitivity and being slightly more efficient (Croux & Dehon, 2010). In the analysis, results are considered significant if the p-value < .05.

Following the correlation analysis, I wanted to apply a logistic regression analysis to test the possibility for students to solve specific numeracy tasks. In this analysis I wanted to examine how reading comprehension affects the possibility of solving tasks, not the latent mathematical ability as measured by UDIR. That is, using the 50 dichotomous variables for each numeracy task as dependent of the ordinal, discreet variable reading comprehension. To make sure that a logistic regression model would yield valid results, I needed to make sure that some properties about my dataset were met.

Firstly, the dependent variable should be measured on a dichotomous scale. As described above, I wanted to measure the probability to get a single numeracy task correctly. As a quick recall, those 50 variables defining the student scores on numeracy tasks are all dichotomous of nature, with 1 for correct and 0 for incorrect answers. Thus, the first assumption is met. Secondly, the predictor variables should be measured on either a continuous or categorical level. In this case, I planned on using reading comprehension as the predictor variable, which is treated as an ordinal variable, though with some scale properties. Thus, all the logistic regressions conducted had a categorical, ordinal predictor variable and a dichotomous dependent variable, and the variables fit the assumptions of the logistic regression model.

Thirdly, there must be an independence of observations. As a researcher, it can be difficult to completely eliminate this source of error, especially in my case where I have not collected the data myself. As described in the chapter *3*, the dataset is a random sample by UDIR of 5854 students from all over Norway, and there are no clusters of data entries (i.e., entire districts being included in the sample, skewing the randomness). Combined with the rigor of preparing and conducting the tests, I found the assumption of independence of observations to be upheld even though I could not directly influence this.

4.4. Ethical considerations

My research, being quantitative, did not have direct contact with the participants, nor did I have any influence in the collection of data. Therefore, most of the ethical reflections regarding those parts of the research is already addressed by UDIR (Utdanningsdirektoratet, 2018a). Still, there are a few ethical considerations to be reflected upon. One of which is how representative the student sampling is. The national tests are obligatory, but it is up to each school to consider exemptions for students based on learning disabilities, although the students need to apply individually for exemption. By omitting the lowest-scoring students, the average score of the school can be artificially high compared to reality. Due to the fact that each school determines the outcomes of the applications for exemptions, when to exempt students is not standardized.

In my methodology, I used the students' reading comprehension raw scores to separate them into low, medium, high, and top performers, with a subsequent correlation analysis to highlight differences. By doing this, I constrained students into four boxes, solely based on their scores on the reading comprehension test. The tests are only indications of the students' ability, so students might over- or underperform on that specific day, which then places them in a box that does not fit their actual ability. Moreover, by compiling students into a box and performing analyses on the data, I generalize the students in each sub-group, while in reality their motivation, ability, and performance might vary greatly. Thus, I recognize that the categorizations made, albeit being necessary to highlight the differences in the student group, is a simplification of the variety found in reality.

5. Descriptive statistics

5.1. Implementation

Before conducting the correlation analysis, I wanted to find some descriptive statistics of the dataset. I used MS Excel to calculate the mean scores of student responses for each task in both reading comprehension and numeracy tests, and subsequently found the mean scores for each category on the reading comprehension test (*find, interpret, and reflect*) and each complexity level on the numeracy test (*complexity level 1-5*). This was compiled into a table, for ease of comparison.

To answer the research sub-question "*how do task properties affect subgroups of students differently, based on their reading comprehension level?"*, it is insufficient to analyze the student base as a single entity, as it lacks the nuance to separate low, medium, high, and top performers in reading comprehension. To battle this, I split reading comprehension scores into 4 categories: 0-11 points, 12-22 points, 23-33 points and 34-44 points, henceforth addressed as low, medium, high, and top performers. It is important to recognize that these categories are not identical to how the results are interpreted in the national tests. As discussed in chapter 3, IRT adjusts the score of each task, making sure that answering correctly on tasks with higher complexity gives more weight to the overall score than answering correctly on the low complexity ones. Thus, students with the same raw score might be placed into different ability levels. However, for this analysis the simplified four group system mentioned above was deemed sufficient and applied.

Furthermore, I wanted to highlight some properties of the student sub-groups based on their reading comprehension ability. The properties included the number of students, gender representation, their average score on the reading comprehension test and if the average score in the group was skewed. This was complemented by a breakdown of how each student group performed in the different categories of the reading comprehension test, and the complexity levels of the numeracy test. The results were discussed, and they made for some assumptions to be tested going into the correlation and logistic regression analyses.

5.2. Results of descriptive statistics and group representations

I split the total student group into four levels of reading comprehension, with information about the gender representation, average scores, and how they did on the different categories and complexities in table 1 and 2 below.

Table 3 - Descriptive statistics of sub-groups

Reading comprehension group	**Number of students**	**Average score reading comprehension**	**Skewness**
Low performers	521	8.50	.772
Medium performers	2049	17.36	.578
High performers	2335	27.89	.535
Top performers	949	36.88	.353

There were a few interesting differences in the student groups to be highlighted. When looking at the number of students in each reading comprehension group, they differ greatly, which was expected. In general, the student group is slightly skewed towards the higher half of the reading comprehension scale, with .56 of the students in the two top levels.

The mean scores also point out a clear trend in the distribution of the data. For the low performers, the mean score is skewed towards the top end of the scale with .77, while the top performers have .35 and are skewed towards the lower end. This makes sense, due to tests including easier questions to differentiate between the low performers, leading to more students being able to solve the easiest questions. Hence, students with low reading comprehension ability are still able to score some points through these entrance questions. Likewise, difficult questions are included to separate the top from the high performers, such that more students would be in the lower end of the top performing scale.

Table 4 - Mean scores of each category and complexity level

	Reading comprehension categories			**Numeracy complexity level**				
Reading comprehension group	**Find**	**Interpret**	**Reflect**	**1**	**2**	**3**	**4**	**5**
Low	.190	.222	.157	.541	.400	.274	.139	.045
Medium	.413	.423	.356	.745	.601	.395	.209	.088
High	.646	.676	.597	.872	.772	.589	.368	.179
Top	.850	.873	.834	.932	.878	.775	.567	.367

Students from all sub-groups score evenly in the find and interpret tasks, with reflect situated slightly lower. Still, the differences are not considered of sufficient size to determine that reflect tasks are harder to solve than the two other categories. As shown in the columns 1-5, every student group follow a similar pattern for the different complexity levels of the numeracy tasks. One particularly interesting result is that in complexity level 5 every group have a mean score about twice as high as the group below them.

The data also supports the notion that students struggle more with numeracy tasks of higher complexity. Also, based on this table it could look like students with high reading comprehension ability generally scores higher than students with low reading comprehension ability, which would be natural to assume. To further investigate this, I conducted a correlation analysis including reading comprehension scores for each student group and the numeracy tasks.

6. Correlation analysis

To assess the relationship between reading comprehension and mathematical word problems, I conducted a Kendall's Tau-b correlation analysis between the ordinal variables "reading comprehension raw score" and "numeracy raw score". This was done with two thoughts in mind; establishing that there was a relationship worth investigating further and comparing the properties of this dataset to similar studies. To make sure that my choice of Kendall's Tau-b did not alter the results in a way that makes it non-comparable to other research, I conducted a Spearman's Rho correlation test on the same two variables. Therefore, I could conclude that the results warranted further investigations and that the dataset looked, initially, to have some similar properties to other studies.

The next step was to examine the correlation coefficients between reading comprehension and the individual numeracy tasks. By analyzing these results, I could evaluate what tasks were having what I considered strong correlation coefficients, that is having a stronger relationship between reading comprehension and the ability to solve the task. Following is a description of the methodology considered in the correlation analysis, with reflections about some of the choices I made. Discussions of the results is also included at the end of this chapter, as I found it to be beneficial to complete the correlation analysis and subsequent discussion before the regression analyses.

6.1. Implementation

Since I conducted the correlation analyses for both the total student groups and the sub-groups based on their reading comprehension level, it was natural to present the implementations of the two separately, as the approaches varied slightly.

6.1.1. Methods of correlation analysis for the total student group

Kendall's Tau-b reports correlation coefficients, which is a measure of strength for the relationship being measured, that is between two variables. An issue arising when dealing with correlation coefficients is how to interpret the magnitude of the coefficients. This is because the thresholds for what defines a correlation coefficient as weak, moderate, or strong differ vastly depending on the area being studied. Some reports show that a correlation coefficient of .300 can be interpreted as weak, moderate or fair according to different fields of study (Akoglu, 2018). Furthermore, others argue that the threshold for significant relationships of .500 is artificially high and that the significance level is depending on what is being measured (Hemphill, 2003). In my analysis, the relationships being studied is two components that are not primarily meant to correlate, i.e., that reading comprehension is not the main component necessary to solve mathematical tasks. Thus, expecting coefficients close to 1, which would mean perfect correlation, is not feasible for this kind of research. In this paper, correlation coefficients are considered weak in the [.000 to .150] range, moderate in [.150 to .300] and strong for [.300 and upwards]. The chosen thresholds may be considered artificial by other fields, but it is important to remember that the correlation coefficients are not a definitive measure of the strength between two variables. Neither do they singlehandedly determine relationships; they are a tool to examine if an increase in one variable could increase or decrease another variable.

I wanted to see which numeracy tasks had the highest correlation with the reading comprehension raw score. The Kendall's Tau-b correlation test was conducted, with results concluded in a correlation table through SPSS. The correlation coefficients were colour-coded in equal intervals such that every numeracy task with a correlation higher than .300, hereby defined as Extracted Numeracy Tasks (ENT), could be extracted.

Furthermore, I was curious as to whether a correlation coefficient from the ENT tasks would be considered significantly stronger than a coefficient excluded from the ENT tasks. Comparing correlation coefficients must be done with caution, with a multi-step procedure that reports if there are statistically significant differences. Walker (2003) built on existing literature to convert Kendall's Tau-b into Pearson's r, such that a Fisher transformation could be applied. This is done using the formula by Kendall, cited by Walker (2003):

$$\boldsymbol{r = \sin(.5\,\pi\,\tau)}$$

Where r is the Pearson coefficient and τ is the Kendall's Tau-b coefficient. This was done for every Kendall's Tau-b coefficient obtained from correlation reading comprehension and the numeracy tasks, such that I had the Pearson coefficients. Next, I had to convert the r into Z_r, using a Fisher transformation.

$$\mathbf{Z}_r = \tfrac{1}{2} \mathbf{ln}_e [(\mathbf{1} + \boldsymbol{r}) / (\mathbf{1} - \boldsymbol{r})]$$

A benefit of the z-value obtained from the Fisher transformation is that it is normally distributed, therefore making comparisons between different z-values easier than for values of Pearson's r (Walker, 2003). After obtaining Z_r, I tested samples from the ENT tasks and the remaining tasks to see if there were significant differences in the correlation coefficients. This was investigated using the Welch's t-test, and I reported both the critical t-value and the level of significance. The reason for choosing the Welch's t-test is that it handles differences in sample sizes and variances well, therefore being suited for the two groups with $N_{ENT} = 8$ and $N_{Rest} = 42$.

Another issue I was faced with in the results is the lack of linguistic dimensions to the numeracy tasks. It was highlighted in the theoretical framework that the existing tools fail to sufficiently analyze word problems, due to the tools being designed for longer texts (Homan et al., 1994). Therefore, when examining specifics of how one task differs from another, I had to be cautious when discussing the linguistic properties. However, I made the decision to loosely discuss some general properties of the numeracy tasks. This was due to the properties, even though not sufficiently anchored theoretically in this thesis, being relatively obvious and subjectable to interpretation. Moreover, I reviewed a sample of the tasks in the light of the cognitive framework by Stein & Smith (1998). As a reader, keep in mind that the parts about linguistic properties of tasks is considered merely a discussion acting as a base for possible future investigations.

6.1.2. Methods of analysis for reading comprehension sub-groups

I previously described how I split the student group into sub-groups based on their reading comprehension ability, such that I had four sub-groups: low, medium, high, and top performers. I wanted to apply the correlation test on the sub-groups to look for varieties in strength of the relationships between reading comprehension and numeracy tasks. To best structure the information, I sorted the numeracy tasks on complexity levels and tested them against reading comprehension raw score, with corresponding correlation tables for each complexity level. Every reading comprehension sub-group were included in each table, so that differences could easily be highlighted and discussed. Lastly, I wanted to explore the ENT tasks for each sub-group, to investigate if these tasks also hold up as having the strong relationships shown in the correlation analysis for the total student group.

6.2. General findings for the total student group

In the correlation analysis testing the relationship between reading comprehension and numeracy raw scores, I received a correlation coefficient of .502 from the Kendall’s Tau-b correlation test, showing that there exists a relationship between reading comprehension and numeracy in the national tests. Furthermore, the Spearman’s Rho correlation gave a coefficient of .677. Thus, even though Kendall’s Tau-b reported a more conservative correlation coefficient, they are both within .47 to .76 which is shown to be normal among researches in the field (Vilenius-Tuohimaa et al., 2008). This merit the investigation of the specific relationships between reading comprehension and each individual numeracy task.

From the correlation test between reading comprehension raw score and numeracy tasks, there were some interesting findings asking for further investigation. Numeracy task 3, 7, 14-16, 18, 24, and 46 all had a correlation of > .300, implying that a strong relationship between reading comprehension and these tasks exist. To investigate whether a random ENT tasks would be considered to have a higher coefficient than a non-ENT task, I converted the tau coefficient to Spearman’s Rho, and subsequently converted the Rho value to a z-value using a fisher transformation (Walker, 2003). By using the Welch’s t-test for two samples with assumed different means, I found that there was a significant difference in the correlation coefficients between ENT tasks and the remaining tasks (t=2.228, p=0.0086), giving ground for discarding the null hypothesis that there are no differences in the correlation coefficients for the two groups. Thus, based on this data I could conclude that the relationship between the ENT tasks and reading comprehension are stronger than for the remaining tasks.

6.3. Correlation results for student sub-groups

The aim of this chapter is to assess how the individual numeracy tasks correlated with reading comprehension, for the low, medium, high, and top performers of reading comprehension, respectively. In both the correlation and the regression analyses, I refer to low, medium, high, and top performers. I would again like to emphasize that this exclusively refers to the sub-groups based on their reading comprehension ability, not to their numeracy ability.

6.3.1. Correlations complexity level 1

Table 5 - Correlations complexity level 1

Reading comprehension group / Numeracy task	**Low**	**Medium**	**High**	**Top**	**Total student group**
Task 1 corr.	.062	.080***	.082***	.050*	.213***
Task 13 corr.	.013	-.003***	.075***	.045	.224***
Task 39 corr.	.075*	.128***	.057***	.035	.248***

* - Significant on a .05 level

** - Significant on a .01 level

*** - Significant on a .005 level

An obvious, first observation is that the groups with fewer students had insignificant results. For the low performers, only task 39 had sufficient data to determine a weak, positive relationship between reading comprehension and the solving of the task. Thus, the data is insufficient to determine that there is a relationship different from zero for task 1 and 13 for this group. As we will see later, few results from the low-performing group were statistically significant. For the medium performers, all results were significant, with task 1 and 39 having weak, positive correlations. Interestingly, task 13 had a marginally negative coefficient. This is in stark contrast to the coefficient for both the other groups and the total student group for this task. The high performers also had significant results on every task with every task now being weak, and positive. For the top performers only task 1 was statistically significant. When comparing to the total student group, there are substantial differences in the correlation coefficients. For the total student group all tasks were significant, and they are all correlating moderately with reading comprehension.

6.3.2. Correlations complexity level 2

Some of the ENT tasks are of complexity 2 and will be discussed in their own sub-chapter. Therefore, they were not included in this table.

Table 6 - Correlations complexity level 2

Reading comprehension group / Numeracy task	**Low**	**Medium**	**High**	**Top**	**Total student group**
Task 2 corr.	.125***	.108***	.081***	.070**	.246***
Task 10 corr.	-.014	.052***	.085***	.036	.240***
Task 27 corr.	.048	.116***	.048***	.021	.218***
Task 29 corr.	.121***	.109***	.047***	.054*	.239***
Task 31 corr.	-.006	.065***	.097***	.055*	.217***
Task 33 corr.	.064*	.123***	.032*	.040	.251***
Task 41 corr.	.041	.116***	.065***	.073***	.226***
Task 42 corr.	.098***	.088***	.090***	.070*	.215***
Task 50 corr.	.122***	.113***	.079***	.065**	.234***

* - Significant on a .05 level

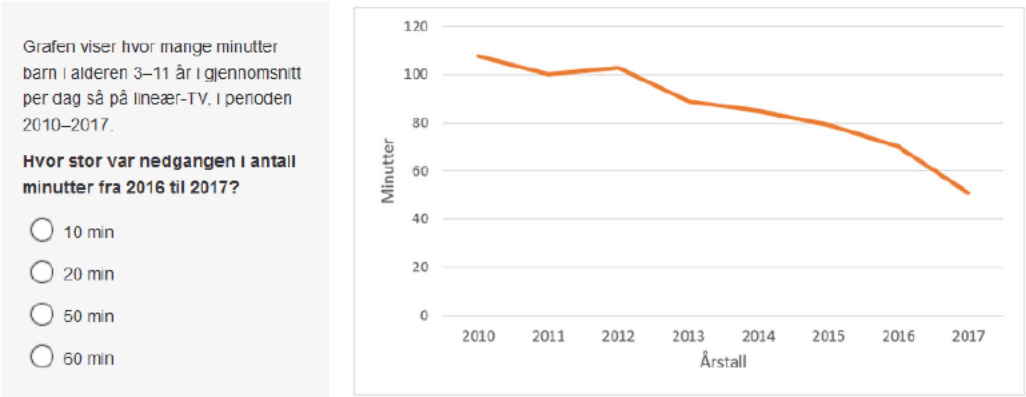
** - Significant on a .01 level

*** - Significant on a .005 level

There were some interesting findings in this complexity level. Tasks 2, 29, 42 and 50 had significant results for all four student groups, and a common feature for these tasks was the decrease in coefficient when moving towards higher reading comprehension performers. Another feature of task 2, 29, 42 and 50 is that they are all higher than the highest ENT correlation coefficient for the low performers. This is an indication that the low-performing reading comprehension students benefit from their reading comprehension on different numeracy tasks than the other groups. In the logistic regression analysis, I will examine the properties of these tasks further. The medium and high performers had significant results for every task, leaving the insignificant results entirely to the low and top performers.

Tasks 31 and 42 had statistically significant coefficients for the top performers, while insignificant for the low performers.

Figure 1 - Task 31



Grafen viser hvor mange minutter barn i alderen 3–11 år i gjennomsnitt per dag så på lineær-TV, i perioden 2010–2017.

Hvor stor var nedgangen i antall minutter fra 2016 til 2017?

○ 10 min

○ 20 min

○ 50 min

○ 60 min

Task 31 aims to measure by UDIR: *Interpret and extract information from graph/table* (Utdanningsdirektoratet, n.d.-a). As discussed in the theoretical framework, interpreting, and extracting information from graphs and visual representations are considered cognitively demanding. Thus, having task 31 being significant for the three upper reading comprehension groups support the struggle in this area of word problems for low-performing students.

There were no correlation coefficients in this complexity level that implied a positive, moderate relationship between reading comprehension and numeracy. Nine coefficients exceeded .100 and they were all between the low and medium performers. Thus, even though the tasks are categorized as having weak relationships, the two lowest-performing reading comprehension groups displayed slightly stronger relationships. This means that the correlation test is more certain in determining relationships with reading comprehension for the two lowest sub-groups, and that their relationships are stronger than for the high- and top-performers of reading comprehension.

Just as for the complexity level 1 tasks, all the tasks correlated moderately with reading comprehension and were statistically significant for the total student group. It is however important to keep in mind that all tasks with correlation coefficients exceeding .300 for the total student group was excluded here. Therefore, even though the tasks included in complexity level 1 and complexity level 2 had approximately equal correlation coefficients, the highest-ranking ones from complexity level 2 was not included in the comparison.

6.3.3. Correlations complexity level 3

Just as for the complexity 2 group, some of the ENT tasks are of complexity 3 and will be discussed in their own sub-chapter.

Table 7 - Correlations complexity level 3

Reading comprehension group / Numeracy task	**Low**	**Medium**	**High**	**Top**	**Total student group**
Task 4 corr.	-.096**	.070***	.114***	.105***	.247***
Task 6 corr.	.055	.079***	.089***	.049*	.205***
Task 8 corr.	.024	.039*	.113***	.072***	.210***
Task 9 corr.	.029	.122***	.105***	.071**	.286***
Task 11 corr.	-.017	.088***	.107***	.120***	.262***
Task 20 corr.	.005	.076***	.108***	.093***	.252***
Task 21 corr.	.049	.091***	.068***	.041	.244***
Task 22 corr.	-.043	.038*	.105***	.125***	.224***

Task 30 corr.	.061	.048**	.071***	.080***	.215***
Task 32 corr.	.057	.051**	.100***	.080***	.228***
Task 35 corr.	.025	.077***	.081***	.070***	.227***
Task 38 corr.	.055	.111***	.123***	.019	.264***
Task 43 corr.	.032	.088***	.087***	.065*	.252***
Task 48 corr.	.104***	.085***	.087***	.054*	.230***

* - Significant on a .05 level

** - Significant on a .01 level

*** - Significant on a .005 level

In the complexity level 3 tasks, only two had significant results for the low performers. Task 4 surprisingly had a statistically significant, negative coefficient for the low performers, contradicting the positive results from the three other groups. Both the medium and high performers had significant, positive results for every task, with tasks 21 and 38 only significant for these two groups.

There are no coefficients indicating moderate correlation between reading comprehension and numeracy besides the ones from the total student group. Compared to the complexity level 2 tasks, there were more coefficients exceeding .100 in this complexity level. They are spread across all student groups, with the majority being in the high performers. Also, 11 out of the 14 tasks had their highest correlation coefficient in the medium or high performers. This supports students with medium-high reading comprehension ability making better use of it when solving higher complexity tasks. This is worth examining further in the next complexity levels.

The correlation coefficients for the total student group shows moderate correlation with reading comprehension, with some tasks (9, 11, and 38) close to the threshold of being categorized as having strong correlation. Similar to the complexity level 2 tasks, there were extracted tasks from this complexity level that were excluded due to them being in the ENT category. Therefore, the coefficients for the total student group discussed here does not paint the full picture. By including the ENT tasks with complexity level 3, it would become evident that both level 2 and level 3 have higher coefficients than level 1 tasks.

6.3.4. Correlations complexity level 4

Table 8 - Correlations complexity level 4

Reading comprehension group / Numeracy task	**Low**	**Medium**	**High**	**Top**	**Total student group**
Task 5 corr.	-.007	.061*	.076***	.105***	.232***
Task 19 corr.	.043	.069***	.108***	.173***	.271***
Task 23 corr.	-.033	.017	.089***	.120***	.214***
Task 25 corr.	.021	.061***	.123***	.137***	.267***
Task 26 corr.	-.058	.120***	.092***	.028	.299***
Task 34 corr.	-.053	-.009	.066***	.064*	.176***
Task 37 corr.	.054	-.002	.118***	.090***	.213***
Task 40 corr.	.020	.092***	.099***	.101***	.278***
Task 45 corr.	.063*	.077***	.125***	.118***	.291***
Task 49 corr.	-.023	.084***	.099***	.098***	.258***

* - Significant on a .05 level

** - Significant on a .01 level

*** - Significant on a .005 level

Only task 45 was significant for the low performers, which is a task measuring *subtraction and multiplication*. Due to no other tasks being significant for this group, the data did not support any conclusions for the low performers, and they will not be discussed further. Another important aspect is the increase of statistically significant results for the top performers. Most of the significant results for this group exceeds .100, and task 19 was the first to correlate moderately for a sub-group. As opposed to the lower complexity levels, the medium performers only have seven out of 10 tasks with significant results. Few of the tasks have coefficients exceeding the .100 threshold. A common feature of these tasks is that the coefficients were low for both the medium performers and the total student group, thus I cannot say whether the insignificant results are due to the performances in the group or due to the tasks not correlating with reading comprehension in general.

All tasks were statistically significant for the high performers, with most of them having coefficients in the upper half of the weak correlation range. Combined with the results for the top performers, the data could imply that reading comprehension favors high to top performers of reading comprehension when solving complexity level 4 tasks.

Tasks 26 and 45 were very close to being included in the ENT tasks, with correlation coefficients of .299 and .291 for the total student group. Thus, even though values in this group are the highest yet, I would like to stress that this does not mean that reading comprehension correlates higher with this complexity level than with the lower ones. We do, however, see that six out of the 10 tasks have correlation coefficients > .250.

6.3.5. Correlations complexity level 5

Table 9 - Correlations complexity level 5

Reading comprehension group / Numeracy task	**Low**	**Medium**	**High**	**Top**	**Total student group**
Task 12 corr.	.081*	-.003	.073***	.044	.172***
Task 17 corr.	.009	.049***	.085***	.108***	.220***
Task 28 corr.	.059	.065***	.110***	.117***	.280***
Task 36 corr.	.052	.007	.082***	.166***	.198***
Task 47 corr.	-.004	.058***	.104***	.133***	.271***

* - Significant on a .05 level

** - Significant on a .01 level

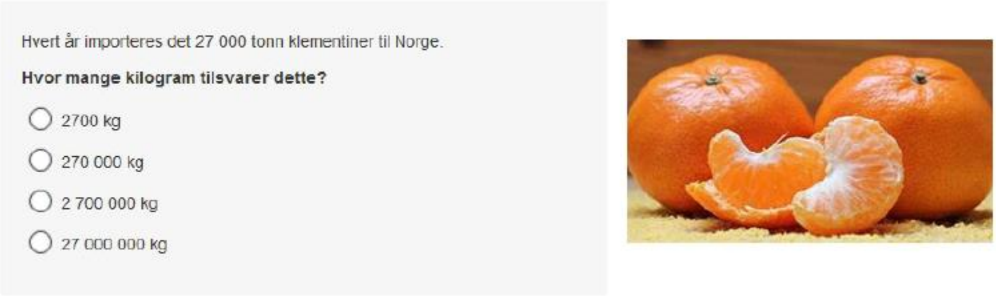
*** - Significant on a .005 level

Complexity level 5 tasks are, as mentioned in the theoretical framework, a tool to differentiate between the high- and top performers. Therefore, it would be unnatural to include an equal number of level 5 as level 3 tasks. Consequently, the dataset for this complexity level might lack some of the nuance that comes with a higher number of tasks.

Task 12 was the only statistically significant task for the low performers, so there are little to no possibility to determine a relationship between the low performers of reading comprehension and complexity level 5 tasks, based on this data. This might be due to them not solving them correctly, as shown with the mean score of .045 for level 5 tasks. The properties of this task will be evaluated in the regression analysis.

The same argument of students not solving enough tasks correctly to get meaningful results can be transferred, to a certain degree, to the medium performers. Data provided support for a positive, weak relationship between reading comprehension and numeracy for three out of five tasks, but the coefficients are considered on the lower half of the weak bracket.

A comparison between the high and top performers shows that the former had a significant result on task 12 while the latter had not. This looks irregular according to the general trend for these tasks, where the top performers in the four other tasks had significant results and exceeded the coefficients for the high performers. One reason for this could be the difference in group sizes, with the top performers having less than half the number of students in the high performers. Another explanation could be in the phrasing of the task:



Hvert år importeres det 27 000 tonn klementiner til Norge.

Hvor mange kilogram tilsvarer dette?

○ 2700 kg

○ 270 000 kg

○ 2 700 000 kg

○ 27 000 000 kg

Figure 2 - Task 12

The length of the sentences is in the shorter end of all the numeracy tasks, and the task is arguably in the intersection between word problems and intra-mathematical tasks with a clear procedure, thus having low cognitive demand. Therefore, failing to determine a relationship between reading comprehension and solving of the task for the top-performers of reading comprehension might be explained by this.

As for the other groups, every task was significant for the total student group. Two out of five tasks were close to the threshold of being extracted as ENT tasks, but every task is classified as correlating moderately with reading comprehension. Task 12 was close to the threshold for weak correlation, with possible reasons for this previously discussed.

6.3.6. *Correlations ENT tasks*

As mentioned in the methods section, the extracted numeracy tasks (ENT tasks) are the ones with correlation coefficients exceeding .300 with reading comprehension. They are originally from complexity levels 2 and 3, with no tasks from the other complexity levels exceeding the threshold. The parenthesis following the task number in table XX represents what complexity level the task was extracted from.

Table 10 - Correlations ENT tasks

Reading comprehension group / Numeracy task	**Low**	**Medium**	**High**	**Top**	**Total student group**
Task 3 (3)	.026	.150***	.171***	.140***	.382***
Task 7 (2)	-.025	.092***	.112***	.134***	.336***
Task 14 (3)	.016	.064***	.169***	.166***	.340***
Task 15 (2)	.025	.130***	.134***	.034	.312***
Task 16 (2)	.080*	.132***	.146***	.117***	.353***
Task 18 (3)	.046	.127***	.151***	.032	.326***
Task 24 (3)	-.026	.125***	.139***	.141***	.352***
Task 46 (3)	.017	.215***	.130***	.135***	.384***

* - Significant on a .05 level

** - Significant on a .01 level

*** - Significant on a .005 level

For the low performers, the ENT tasks were, except for task 16, statistically insignificant. Every task had significant results for the medium and the high performers, while six out of eight were statistically significant for the top performers. All tasks correlated strongly with reading comprehension, which was a chosen threshold for extraction.

Moreover, there were several tasks that were classified as correlating moderately with reading comprehension for the sub-groups. Task 3 correlated moderately for both the medium and low performers, with the low performers being close to moderate. Task 14 correlated moderately with the high and top performers. Task 18 and 46 correlated moderately with the high and medium performers, respectively. All tasks with moderate correlation coefficient are from complexity level 3, and there are six moderate coefficients in total. Compared to two moderately strong coefficients for the remaining 42 tasks across all student sub-groups, this is a clear indication of the ENT tasks possessing a stronger relationship with reading comprehension than the other tasks.

When examining the results, I proposed some ideas based on some general properties of the tasks, which could act as bases for further research. This is especially important for the ENT tasks, as the impact of identifying properties could deepen our understanding of how to teach and distinguish between different types of word problems.

As mentioned in the general findings, there are no indications that specific types of input format in the tasks affect the correlation coefficients. This is also applicable here, with the ENT tasks having multiple choice, composite, and numeric answer options, with none displaying higher correlation coefficients than the others. A common denominator for all ENT tasks is that they are all considered to have high cognitive demand.

6.4. General discussions of correlation results

To summarize the findings from the correlation analysis, there were some interesting results that helped answering my research question. Even though the sub-group sizes varied greatly, which can affect the results, some tendencies among the different reading comprehension groups were found. The low performers were the group with fewest students, and they only had four tasks from complexity 3-5 with significant results. This might be due to them not being able to solve enough tasks on these complexity levels to determine a relationship with reading comprehension. As shown in *table 4*, they had an average score of .139 for complexity 4 tasks and .045 for complexity 5 tasks. This means that if the students solve almost no tasks correctly, it is near impossible to establish whether there is a relationship with reading comprehension for said tasks. Thus, based on this analysis it seems like students with low reading comprehension ability benefit most from said ability when solving low complexity mathematical word problems.

Another important aspect was the difference between the high and top performers in complexity level 4 and 5 tasks. This is an example of how a smaller group does not necessarily have to equal less significant results. The high performers were over twice as many as the top performers, however the latter had more tasks with significant results than the former for both highest complexity levels. Even more, albeit all being categorized as weakly correlating, the top performers had higher coefficients for almost every task in these complexity levels. This result implies that when the complexity level of word problems increase, the students with higher reading comprehension ability makes better use of said ability than the lower performing reading comprehension students.

Moreover, the ENT tasks shone light on some differences between the low performers and the others. As discussed in *6.3.6* just one of the eight ENT tasks had significant results for the lowest performing reading comprehension group, while the other groups had significant results for a majority of the tasks. Combine this with the fact that there were seven tasks with higher correlation coefficients than the highest ENT task for the low performers, there is a clear indication of this low-performing reading comprehension students benefiting from their reading comprehension ability on different kinds of word problems. It is interesting to note that six out of seven tasks are from complexity level 2 and 3, the same levels that the ENT tasks consist of. Therefore, it looks like the tasks that the low performers benefit most from is not necessarily easier than the tasks the other groups benefit most from (the ENT tasks). The difference in the high correlation coefficient tasks for the low performers compared to the high correlation coefficient tasks for the other groups could therefore be a research area worth investigating, and I will address that in the next chapter.

7. Regression analysis

After some discussion of the results in my correlation analysis, there were some questions raised that merit further investigations. Firstly, I found that low performers benefit from their reading comprehension ability on different tasks than the other groups. I addressed this by making logistic regression models for the complexity level 1 and 2 tasks, and subsequently analyzed some properties of the tasks with high correlation coefficients for the low performers. Secondly, I found that for the top performers, the data supported a relationship between reading comprehension and numeracy tasks increasing, when the complexity of the numeracy tasks increased. I investigated this by making regression models of the complexity level 5 tasks, and again discussed some properties of the tasks in light of the cognitive framework by Stein & Smith (1998). Thirdly, the ENT tasks appeared to have some properties separating them from tasks on the same complexity level, and that they had the strongest relationships with reading comprehension. This was also examined through a logistic regression analysis, with the results discussed, giving a total of 25 regression models.

7.1. Implementation of logistic regression analysis

Due to technical limitations in SPSS, the logistic regression analysis was conducted 25 times – each time with a single numeracy task as a dependent variable being subjected by the predictor variable reading comprehension raw score. For the dependent variables, they were coded such that 1, the event happening, was the student solving the task.

There are some aspects of reliability and validity that should be discussed and evaluated when conducting a logistic regression analysis. I based my structure on focus points discussed in the article by Peng et al. (2002) which includes statistical tests of individual predictors, goodness-of-fit statistics, and validations for predicted probabilities (Peng et al., 2002).

Firstly, the statistical test of individual predictors is an indication of whether the model including the predictor would correctly predict student responses with higher accuracy than the intercept model. This is measured using the Wald test, reported by SPSS. The null hypothesis in the Wald test assumes that there are no differences between the intercept model and the model including the predictor. Therefore, a statistically significant results means that the null hypothesis can be rejected, and that there exists a difference between the two models. Thus, the Wald test helped me assess whether reading comprehension would indeed affect the solving of the numeracy tasks.

Secondly, the goodness-of-fit statistics assesses how well the model fits the data. A combination of the Omnibus tests of model coefficients and the Nagelkerke R square was used for my consideration of fit, both reported in SPSS. Nagelkerke R square is considered a pseudo-R squared, that is an attempted equivalent to the explanatory power R^2 from linear regression. This is not to say that the Nagelkerke R squared measures the explanatory power – it is merely an indication of how well the model measures what it aims to do and is used to compare similar models. It is important to note that studies with more in-depth analysis might consider this assessment of goodness-of-fit a bit superficial, but other methods of measurement are outside the mathematical knowledge possessed by me and were therefore omitted.

Thirdly, there should be a validation of the predicted probabilities. I focused on the classification table provided by SPSS, that shows how many percent of the students that the model would predict correctly. The classification table evaluates both the percentage predicted of events, that is students solving the task based on the coding of the variables, and the predicted percentage of nonevents. Though it is not an exhaustive measure of prediction validation, it is considered a measure in the article by Peng et al. (2010) and I deemed it sufficient for this analysis.

Following the discussion about the aspects to be considered in logistic regression, there are several essential data reported in the results that should be mentioned. The discrimination *a*, that is the innate property of the task to discriminate between low and high performers in numeracy, is reported in the “Variables in the equation” table in Block 1 in the SPSS output. Paired with *a* is the constant for the equation, which I arbitrarily chose to name *C*. The discrimination component *a* and the constant *C* can be used to compute the task difficulty *b*. The task difficulty *b* is a measure of the level of reading comprehension needed for having the probability of .500 to solve the task.

The probability of solving a task given reading comprehension as a predictor can be presented as a graph, which is more intuitive than plainly reading the numbers from the tables. The graph below shows the comparison of task 14 and 18, with the x-axis showing the level of reading comprehension and the y-axis being the probability of solving the task. The line connecting the points of both graphs indicate the level of reading comprehension needed for having a probability of .500 of solving the task.

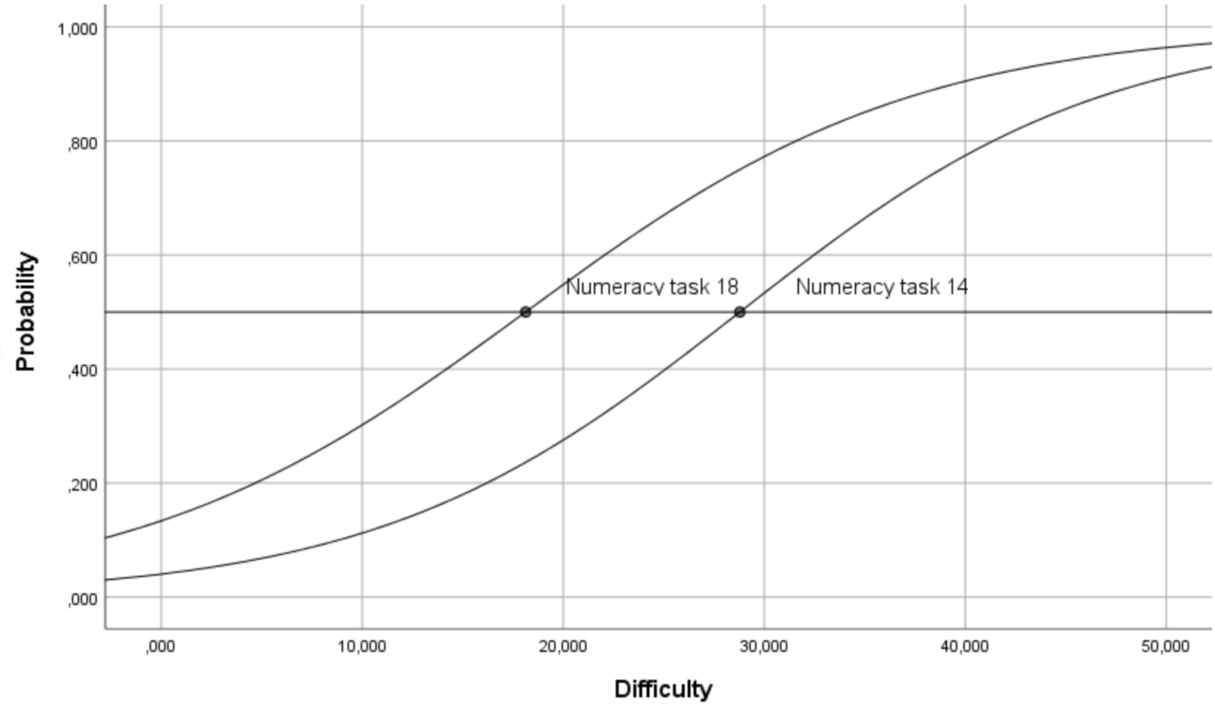


Figure 3 – Regression model graphs of task 14 and 18

This representation highlights how two tasks from the same complexity level, and with nearly equal discrimination coefficients, can still have vastly different difficulty thresholds. The level of reading comprehension required for having a .500 probability of solving task 14 is significantly higher than task 18, showing that the complexity levels insufficiently capture task nuances.

7.2. Regression results for mathematical complexity level tasks

With the generalized aspects of the logistic regression results explained, it is time to delve into the specific results from my dataset. As I explained in chapter *4.4* regarding analysis strategy, there were three areas I wanted to investigate properties of tasks through the logistic regression models: The ones that correlated strongest with reading comprehension for the low performers, the highest complexity ones for the top performers, and the ENT tasks.

7.2.1. Regression results for complexity level 1

Before analyzing the tasks difficulties, I wanted to discuss the assessment results described in the methods section by Peng et al. (2002). When assessing the predictor, that is reading comprehension, the Wald test determines the probability that the model including the predictor is more accurate than without the predictor. For all tasks of complexity level 1 the Wald test had significant results, indicating that the regression model fits better than without the predictor. Nagelkerke R squared ranged from [.104 - .165], showing that the models for the different tasks did not differ greatly. The Chi-squares reported in the Omnibus test of model coefficients show that they were all significant, with values in the range [400 – 590]. This does not explain much now, but I will make a comment about the Chi-squares reported from the ENT task models later in the analysis that will clarify.

The last result I wanted to examine before discussing the regression models is the percentage estimated correctly. More precisely, I aimed to seek out differences in the percentage estimated in the intercept model and in the model including the predictor. The intercept model calculates the average probability of solving the task for all students, and if the probability exceeds .500 it estimates that every student got it right. Conversely, if the overall probability is less than .500 it will guess that every student answered wrong.

There were no significant differences in correctly estimated percentages for the intercept and predictor model for either of the tasks. The intercept model predicted marginally better than the predictor model for task 1 (.791 against .789) while the two others were marginally higher for the predictor model. The overall correct estimations for the predictor models were between .789 and .849. Moreover, the predictor model had correctly estimated close to every event, that is the student solving correctly. However, the percentage estimated of wrong answers were close to zero (.014 to .038), showing that the model estimated students solving correctly better than those who are not able to solve the task. Thus, the models were better calibrated for sensitivity, that is correctly classified events, than specificity (correctly classified nonevents).

To summarize, the predictor models for complexity 1 tasks had some indications of being a better fit than the interceptor models, like the Wald test. However, there were also some aspects that spoke against their significance over the intercept models, namely the lack of differences in percentage estimated correctly between the intercept and predictor models.

Table 11 - Regression results complexity level 1 tasks

Numeracy task	**Corr. With reading comprehension**	**Difficulty threshold (b)**	**Discrimination (a)**
1	.213	4.78	.076
13	.224	6.16	.080
39	.248	5.20	.108

As mentioned in the methods section I did not separate the student group into the reading comprehension sub-groups for the logistic regression analysis. For a quick summary of the correlation analysis, I included the correlation for each task with reading comprehension, and all tasks from complexity level 1 correlated moderately.

The task difficulties were closely related, with all three tasks showing that the necessary reading comprehension skill for having a .500 probability of solving the task were low. This means that the students should be able to solve these tasks, almost no matter their reading comprehension ability. Thus, the complexity one tasks works as intended by UDIR, as having a low threshold for solving and easily accessible for most students. Even though none require significant reading comprehension ability, the slopes *a* differ. Task 39 was affected the most by reading comprehension in the probability of solving the task.

The tasks in complexity level 1 had similar difficulty thresholds, which is an indication that they act as the easily accessible tasks they were designed for. The slopes differed, showing that an increase in reading comprehension does not mean equal increases in probability of solving the tasks. The task with the steepest slope, task 39, also had the strongest correlation coefficient.

7.2.2. Regression results for complexity level 2

I excluded the ENT tasks from complexity level 2 (numeracy task 7, 15, and 16), due to them being discussed in sub-chapter *7.2.4*. Coherently to *7.2.1*, I wanted to address the assessment results of the models by (Peng et al., 2002). Just as for the complexity level 1 tasks, all models gave significant results on the Wald test. Moreover, the Nagelkerke R-squared had values in [.092 to .159], indicating that there were no large discrepancies between the different task models, albeit marginally lower than for the complexity level 1 tasks. The Chi-squares reported were all significant, in [412 – 582]. As I explained in the previous sub-chapter, the values of the Chi-squares form a base for comparison between these and the ENT tasks.

When examining the percentage each model correctly estimated, the increase in percentage was marginally positive for every task. This speaks to the models not being substantially better at predicting the student scores than if reading comprehension was not included, but better than for the previous tasks. The predictor models correctly estimated most of student answers [.832 to .996], but this can be due to the model over-estimating how many correctly solve the task. In a situation where most students gets the correct answer, so the probability threshold exceeds .500, the intercept model would correctly predict most students solving correctly, having a high sensitivity. Conversely, if the probability of solving the task is low, the specificity of the model would be increasingly accurate. Because of the complexity level it is natural that the sensitivities are more precise than the specificities. To strengthen this, the percentage estimates should be addressed in the complexity level 5 tasks, where the specificity should be more accurate than the sensitivity.

Some assessments spoke of the predictor models as being better than the intercept models, like the Wald test and the Chi-squares were significant. However, the percentages correctly estimated revealed that the prediction accuracy only marginally improved over the intercept models. Thus, for these tasks reading comprehension as a sole predictor does not bear a lot of magnitude, but it still gives small indications that are worth examining.

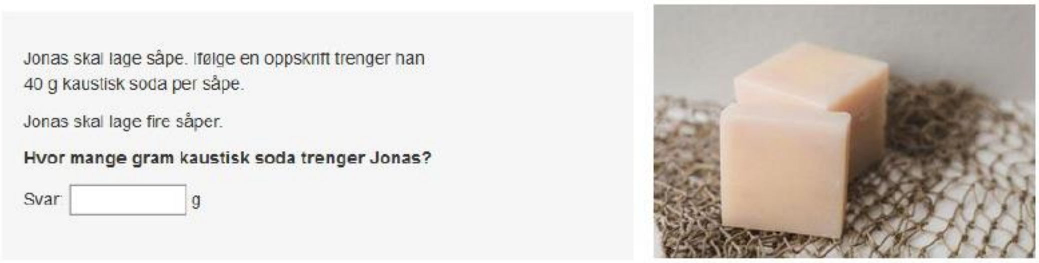
Table 12 - Regression results complexity level 2 tasks

Numeracy task	**Corr. With reading comprehension**	**Difficulty threshold (b)**	**Discrimination (a)**
2	.246	8.81	.086
10	.240	11.89	.075
27	.218	12.04	.067

29	.239	4.65	.106
31	.217	16.28	.064
33	.251	11.19	.084
41	.226	9.85	.074
42	.215	10.01	.069
50	.234	8.90	.080

Firstly, all the tasks included here were moderately correlated with reading comprehension, due to the extracted ENT tasks. The difficulty thresholds had a wider spread than the complexity level 1 tasks, ranging from 4.65 for task 29 to 16.28 for task 31. The spread could simply be due to this complexity level having more tasks, so that a similar pattern might be found if there were more tasks of complexity level 1. However, this data supports a wide spread of difficulty thresholds on the complexity level 2 tasks, implying that there are some innate properties of certain tasks that makes them more or less accessible for low performers of reading comprehension.

By comparison, the properties of task 29 and 31 varied greatly and aimed to measure different skills. I elucidated in the theory that word problems can have varying cognitive demand, and that is evident here. Based on (Utdanningsdirektoratet, n.d.-a) task 29 aims to *measure multiplication of whole numbers*, and students are asked to fill out the correct answer. I have established that there is a distinction between intra-mathematical tasks and word problems, and tasks like 29 is situated somewhere in the threshold between the two:



Jonas skal lage såpe. Ifølge en oppskrift trenger han 40 g kaustisk soda per såpe.

Jonas skal lage fire såper.

Hvor mange gram kaustisk soda trenger Jonas?

Svar □ g

Figure 4 - Task 29

There exists one procedure for solving the task, which is to apply a schema for multiplication, hence this task should be considered as having a low cognitive demand. Contrarily, task 31 asks the student to *extract and interpret the graph*. By asking to interpret, the task invites students to build on their pre-existing knowledge of graphs, applying a strategy to identify an approximate answer and consider the validity of the response. Moreover, the task is multiple choice, giving student the opportunity to dismiss obviously wrong answers. Thus, task 31 possess requirements of high cognitive demand.

In this comparison, task 29 also had the lowest difficulty threshold of the complexity level 2 tasks, while task 31 had the highest. Another aspect of task 29 was the significance levels of the correlation analysis. By recalling to chapter 6.3.2, low performers had statistically significant correlation between reading comprehension and task 29 at a .001 level, while it was only significant at a .05 level for the top-performers. Thus, the data are more confident in determining a relationship with reading comprehension and task 29 for the low performing students. Although I have to be careful in extrapolating the results generally, it does speak to how the difficulty threshold is lower for tasks with low cognitive demand, and that the low performing reading comprehension students benefit from their reading comprehension ability on tasks with low cognitive demand. This would be too comprehensive to investigate in my thesis but makes for a potential implication to be addressed in future studies.

Secondly, the discrimination slopes showed that a unit increase in reading comprehension would yield slightly different increases on probability of solving the tasks. Again task 29 stood out as the only task with a discrimination over .100, meaning that this is the complexity level 2 task where student benefit the most from a potential increase in their reading comprehension ability. This could be explained by the linguistic properties of the task combined with the low cognitive demand of the task. It could also be explained by the low difficulty threshold, or a combination of the two.

To summarize, the complexity level 2 tasks had a wider spread in the difficulty thresholds, showing that the amount of reading comprehension required for having a .500 probability of solving the task varied greatly. I proposed a potential hypothesis in that tasks with low cognitive demand could have a lower difficulty threshold than tasks with high cognitive demand, and that the low performers benefit most from their reading comprehension ability on tasks with low cognitive demand. However, this needs to be investigated further before any certain claims are made. Moreover, the varieties in difficulty thresholds indicated that the tasks aim to measure a wider range of student ability than the complexity level 1 task. Some tasks, like task 29, are easily accessible and fairly straight-forward, inviting all students no matter their skill. Other tasks, however, are more cognitively demanding, where students need to show profound mathematical skill. With an increase in reading comprehension, the probability of solving a task increased for every task in this complexity level, with perhaps a weak favour towards reading comprehension affecting accessible tasks more.

7.2.3. Regression results for complexity level 5

The correlation analysis on complexity level 5 tasks indicated that top-performers of reading comprehension benefits more from said ability when the complexity level of numeracy tasks increases. Therefore, I wanted to include these tasks in the regression analysis, to examine differences and possibly some properties of the tasks.

Similar to the other groups, the Wald test was significant for all five tasks, however there was a larger spread in the values. The Nagelkerke R-squared interestingly also had a wider spread than the other groups [.065 - .215]. Task 12 had the lowest reading comprehension explanatory power of all the tasks analyzed, while 36 and 47 were two of the highest. Compared to the lower difficulty levels, the differences in reading comprehension as a predictor for the tasks became clearer here. Looking at the Chi-square values, the spread was also wider here [262 – 723], with values exceeding the lower complexity levels on both ends of the scale. Still, they were low compared to the values in the ENT tasks, which will be discussed in the next sub-chapter.

I hypothesized previously that the specificity, that is the models correctly estimating the students incorrectly solving a task, would be better calibrated than the sensitivity for the complexity level 5 tasks. This turned out to be true, as the test estimated close to every incorrect answer correctly. Conversely to the accurate sensitivity of the models in complexity level 1 and 2, the models including the predictor did not manage to correctly estimate a sufficient number of cases. Task 28 had the most accurate sensitivity, while also having the lowest difficulty threshold. Task 36, with the highest difficulty threshold, also had the lowest sensitivity calibration. Finally, the predictor models including reading comprehension marginally improved the estimations of the intercept models, but a significant improvement could not be determined.

Table 13 - Regression results complexity level 5 tasks

Numeracy task	**Corr. With reading comprehension**	**Difficulty threshold (b)**	**Discrimination (a)**
12	.172	43.82	.057
17	.220	43.84	.089
28	.280	37.56	.105
36	.198	46.21	.175
47	.271	41.51	.140

The trend of wider spread in the results for complexity level 5 tasks continued with the correlation coefficients from previously. Tasks 12 and 36 were almost considered to have weak correlations [.000 - .150] with reading comprehension, while task 28 and 47 approximated the threshold of being included in the ENT tasks. The difficulty threshold of task 28 was abnormal, situated lower than the other tasks. Still, the models estimated that for solving the complexity level 5 tasks, the student need a score of 40+ on their reading comprehension. This supports the notion that there is a relationship between top-performers in reading comprehension and numeracy.

One surprising result asking for revision was the comparison of task 12 and 36. The Nagelkerke R-squared showed that task 12 had .065 while task 36 had .213, showing that reading comprehension was vastly more important in predicting the outcome of task 36. However, a contradicting argument to this is the correlation analysis, showing that they were both situated on the lower end of moderate correlation [.150 to .300]. Neither task strongly correlated with reading comprehension and were some of the lowest of all the 50 tasks. The discrimination of each task is in support of the first argument, where task 12 had a very low *a* while task 36 had one of the highest of all tasks. Thus, the relationship with reading comprehension seems closer for task 36 than 12.

One way of addressing this is to investigate how UDIR defined the properties of each task. Task 12 aims to measure the *ability to convert between tons and kilograms,* and it is a multiple-choice task. Conferring to the theoretical background, the task is considered a “dressed up word problem” with a pre-defined path of solving, namely the algorithmic process of converting tons to kilograms. It is not considered intra-mathematical due to the context of fruits, but it, similarly to task 29 in complexity level 2, lies in the threshold between the two. Thus, it should be considered a task with low cognitive demand. Task 36, on the other hand, aims to measure the *understanding of mean values*. Again, the word *understanding* implicates that students need to make considerations about the answers and reflect upon it, which is a property of tasks with high cognitive demand. Therefore, it makes sense that a purely textual task, without multiple-choice, are more strongly influenced by the students reading comprehension ability, hence the high *a* for task 36.

As opposed to the discussion between task 29 and 31 in complexity level 2, the difficulty threshold is just marginally different for task 12 and 36. This indicates that there were not necessarily tasks with high cognitive demand that have the higher difficulties, and that there are other properties of tasks than reading comprehension that determines the difficulty. I do recognize that in a study investigating all properties of mathematical word problems, reading comprehension is merely one of many factors. However, due to the scope of my thesis, I chose not to evaluate how the other properties determine the task difficulty. Hence, I let the comparison between task 12 and 36 be an example of how tasks analyzed through my strategy might seem similar, but the similarity stems from completely different backgrounds.

7.2.4. Regression results ENT tasks

The correlation analysis revealed that eight numeracy tasks had properties giving them high correlations [.300 - →] with reading comprehension, so after addressing them separately in chapter *6* it was natural to discuss them exclusively in the regression analysis as well. In the table below, the complexity level of each task is indicated with the parentheses after the task number.

The Wald test revealed significant results for all models, indicating that the predictor models improve over the intercept ones. A comparison of the Nagelkerke R-squared values now become relevant, as complexity levels 1, 2 and 5 had values in the [.065 - .215] range. For the ENT models, the values lay in the range [.193 - .275], such that the ENT model with the lowest explanatory power, task 15, had value close to the highest non-ENT model. This is an indication that reading comprehension has a higher impact on correctly estimating student answers for the ENT tasks.

In the sub-chapters above, I presented the values of Chi-square reported in the regression results. They were situated in the range of approximately [200 – 750] with the majority between 400 and 500. This is in stark contrast to the Chi-squares reported in the ENT models, in the range [878 – 1374]. Again, the results point towards reading comprehension being a stronger indication of the students probability of solving the ENT tasks.

In addition, the percentage correctly estimated is worth investigating. Compared to the models discussed above, the intercept model, that is without reading comprehension as a predictor, have a lower accuracy with .50 to .70 percent correctly estimated. However, when including reading comprehension as a predictor, the accuracies improved with .04 to .16. This indicates that these tasks, that correlated strongly with reading comprehension, also have regression models that predicts the outcome better when including the predictor. Even though the models still did not have perfect accuracies, it should be recognized the significant improvement in some of the tasks. For task 7, the percent correctly estimated improved from .508 in the intercept model to .677 in the predictor model. Moreover, the ENT models possessed similar accuracies of sensitivity and specificity, even though in total being slightly lower than the other models. This might be due to the choice of complexity levels to investigate; by choosing complexity levels 1, 2, and 5, I investigated the tasks that either most of the student correctly solved, or most student incorrectly solved. Thus, the models in the complexity level tasks automatically estimated high percentages correctly in the intercept models. If I examined the complexity level 3 and 4 tasks as well, I might have gotten results with weaker intercept models.

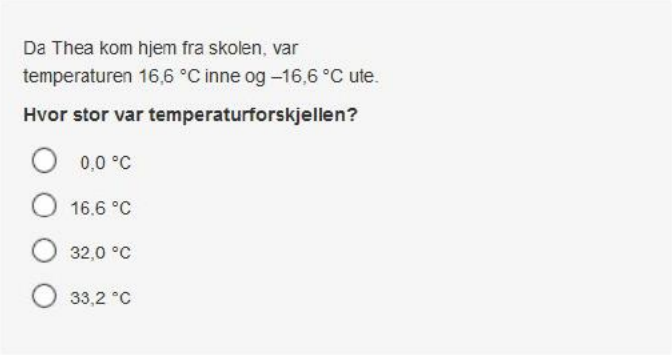
Based on the discussion above, there are several results pointing towards the ENT tasks possessing properties that distinguish them from the other tasks. The Nagelkerke R-squares reported showed that the explanatory power of these models out-perform the non-ENT models, and the Chi-Squares support that notion with the value increase compared to the other models. Moreover, the ENT models had significant improvements over the intercept models, which should be interpreted as an indication of reading comprehension estimating their solving probability better than for the non-ENT models. Even though the intercept models correctly estimated a lower percentage than for the other tasks, that is not really important for the evaluation. What matters is the difference between the intercept and the predictor model, and it becomes clear that the ENT models vastly out-perform the others. Thus, these models yielded relevant results for my research question, and they support the result from the correlation analysis of being closely related to reading comprehension.

Table 14 - Regression results ENT tasks

Numeracy task	Corr. With reading comprehension	Difficulty threshold (b)	Discrimination (a)
7 (2)	.336	24.35	.105
15 (2)	.312	16.10	.101
16 (2)	.353	15.28	.123
3 (3)	.382	25.09	.125
14 (3)	.340	28.79	.110
18 (3)	.326	18.13	.103
24 (3)	.352	27.35	.113
46 (3)	.384	21.74	.127

The correlation coefficients, naturally, all exceeded .300 and were considered to correlate highly with reading comprehension. There were no significant differences in coefficients between the complexity level 2 and complexity level 3 ENT tasks. For both complexity levels of ENT tasks, the difficulty thresholds were higher than the non-ENT tasks of the same complexity. This is specifically evident in the complexity level 2 tasks, where most of the non-ENT tasks had difficulty thresholds around 10. When task 7, 15, and 16 exhibited difficulty thresholds closer to the average of complexity level 3 tasks, it shows that the relationship between reading comprehension difficulty and mathematical complexity is non-linear.

I previously compared tasks with high and low difficulty thresholds in relation to the level of cognitive demand, and that comparison yielded results for this taskset as well. Task 18 has a low difficulty threshold compared to the other complexity level 3 ENT tasks, and by examining the properties of the task, UDIR defined it to measure *negative integers, multiplication.* Examination of the task phrasing showed that the task should not be considered open-started nor open-ended.



Da Thea kom hjem fra skolen, var temperaturen 16,6 °C inne og –16,6 °C ute.

Hvor stor var temperaturforskjellen?

○ 0,0 °C

○ 16,6 °C

○ 32,0 °C

○ 33,2 °C

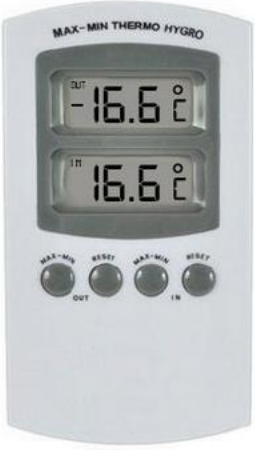
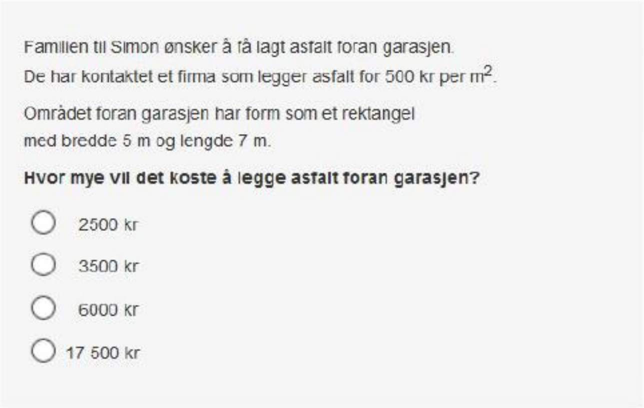


Figure 5 - Task 18

Students can evaluate the validity of their result, but overall, the task seems to have a low cognitive demand due to the procedural nature of solving it.

Task 14 on the other hand, with the highest difficulty threshold, aim to measure *area, multiplication.* Based exclusively on this, task 14 also seems to have a low cognitive demand. However, when looking at the task phrasing there is a lot of information presented:



Familien til Simon ønsker å få lagt asfalt foran garasjen.
De har kontaktet et firma som legger asfalt for 500 kr per m^2.

Området foran garasjen har form som et rektangel med bredde 5 m og lengde 7 m.

Hvor mye vil det koste å legge asfalt foran garasjen?

○ 2500 kr

○ 3500 kr

○ 6000 kr

○ 17 500 kr

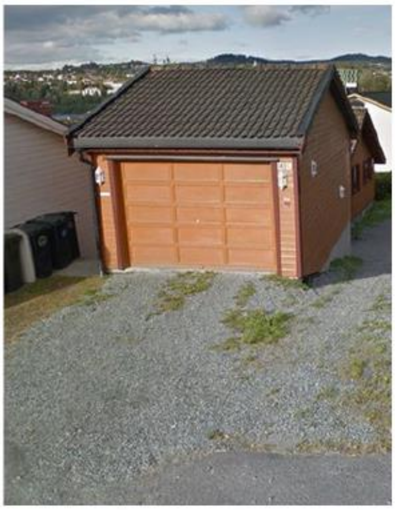


Figure 6 - Task 14

This could also be considered a restricted modelling task – the simplification of the real world into the mathematical is already done by assuming that the area in front of the garage is a perfect rectangle, which contradicts the representation of the area in the picture provided. So, even though task 14 and 18 both have straight-forward procedures for solving, task 14 requires students to connect several mathematical properties, thus having a higher cognitive demand. When looking at task 24, with the second highest difficulty threshold, UDIR aims it to measure *interpret and extract information from a table.* This kind of task, as I reflected on previously, is considered to have a high cognitive demand. Therefore, the comparison of task 14, 18, and 24 showed that the cognitive demand of tasks could help explain the discrepancies of the difficulty thresholds. It could be that the sheer amount of text in task 14 is what makes the difficulty threshold so high, and that is a notion asking for further research.

In general, the logistic regression results from the ENT tasks complemented the findings from the correlational analysis. The model fit considerations showed that reading comprehension as an explanatory power for the probability of solving a task was significantly stronger for the ENT tasks, and this was supported by the improvement of the predictor models over the intercept models. Moreover, the ENT tasks had higher discrimination values than non-ENT tasks, and an increase in their reading comprehension ability more strongly affected the probability of solving the task. Thus, both the correlation analysis and the logistic regression models support the ENT tasks as having properties tying them closer to reading comprehension than the remaining tasks. I briefly reflected upon properties of a few tasks, with a disclaimer that it is by no mean an exhaustive analysis. I discussed difficulty thresholds in the light of the cognitive demand framework and found that it could explain the differences. This conclude the analyses, and I believe that I have identified some properties of word problems, and student sub-groups, that can both build on the existing literature and merit further investigations.

8. Conclusion

In the beginning of my thesis, I aimed to *identify relations between reading comprehension and mathematical word problems of varying complexity,* within the context of the national tests conducted in the autumn of 2020 in Norway. I found in the correlation analysis that the degree of strength in the relationships varied both between and inside the complexity levels. The tasks with correlation coefficients exceeding .300 originated from complexity level 2 and 3, and both analyses showed that they possessed some properties that made them closely tied to reading comprehension. This was further supported in the regression models, where all ENT tasks possessed high difficulty thresholds compared to their original complexity level. Also, an increase in the students reading comprehension ability affected the probability of solving an ENT task more than non-ENT tasks. Therefore, in-depth studies that could identify the properties of ENT tasks could enhance our understanding of word problems, and maybe create a base for differentiating the approach for word problems, based on the properties of the relevant task.

I built on the theoretical background by proposing an idea the level of cognitive demand could explain the difficulty thresholds, supported by comparisons in the logistic regression results. By examining the tasks with low and high difficulty thresholds in each sub-chapter, I did identify that the tasks that was considered low cognitively demanding tended to have low difficulty threshold. Conversely, I found that the highest difficulty thresholds seemed to have high cognitive demands. In regard to my research question, I found that there were differences for tasks in their relationship to reading comprehension, inside the same complexity levels. Thus, my analyses imply that evaluating the students' numeracy ability through solved tasks on each complexity level fails to recognize the nuances in task difficulties. Therefore, considerations to the cognitive demand would deepen the understanding of each task, subsequently being more precise in evaluating what each student is lacking, and mastering.

Also, I wanted to investigate if *the task properties affected student sub-groups differently.* I discovered that the low-performers seemed to benefit most from their reading comprehension ability on other tasks than the others. This was evident in task 2, 29, 41, and 50 where the low-performers had the highest correlation coefficients, even though being the sub-group with fewest students. Moreover, for the low-performers, these tasks had higher coefficients than the ENT tasks, showing that the ENT tasks was not the most influential for this sub-group.

Further support was provided by the regression models, where task 2, 29, 41, and 50 possessed the lowest difficulty thresholds of all the complexity level 2 tasks, just slightly higher than the complexity level 1 tasks. All this showed that students who struggle with reading might need help with other kinds of tasks than students better at reading, and further investigation into this could help low-performers with identifying problematic task structures, and conversely build their confidence on specific tasks.

For the top-performers, the correlation analysis could not determine relationships between reading comprehension and the lower complexity level tasks. However, as the complexity level of the tasks increased, so did the number of significant results, ultimately out-performing the other groups on the highest complexity levels. This is an interesting result, because the preliminary tests revealed that top-performers had a mean score of .932 on the complexity level 1 tasks. Even though almost all the top-performers of reading comprehension managed to solve the lower complexity tasks, the data could not determine a relationship between their reading comprehension and the given word problems. The sample size of the top performers is sufficiently large to avoid being a cause, so there might be other factors in play.

For the highest complexity level, I discussed why tasks 12 and 36 differed in their difficulty threshold, and why the high performers had significant correlation result for task 12 while the top performers did not. Task 12 was considered to have a low cognitive demand, while task 36 was considered to have a high demand, showing that the complexity levels fail to fully explain the task difficulty. That is supported by the difficulty thresholds, with task 12 being significantly lower than task 36. A general proposition from this is that students who have a low probability of solving a task given their reading comprehension level, are more likely to solve the task if the cognitive demand is low.

To summarize, I found *several relations between reading comprehension and numeracy tasks of varying complexities.* The tasks confined in one specific complexity level should not be treated equally, as the amount of reading comprehension required, the task difficulty and the cognitive demand of each task varies. Moreover, by *examining the task properties for different subgroups based on their reading comprehension level,* I found that low performers benefit most from their reading comprehension ability on low difficulty tasks with low cognitive demand, and those tasks are not necessarily confined to complexity level 1. Moreover, the top performers showed that as the task complexity increased, so did the relationship with reading comprehension.

8.1. Limitations

As with most studies, there are several limitations that has been considered throughout the process. Some limitations are already discussed in their relevant chapter, so this sub-chapter structures the notion previously discussed while simultaneously introducing new aspects.

Because of my choice of a quantitative research design, I missed the opportunity to conduct interviews with students. Previous researchers (G. Nortvedt, 2009) found interesting results by discussing task strategies with the students, and the qualitative data from such a design could have complemented the results from the national tests. Moreover, I did not have the means to distribute and execute tests based on my own research, thus creating an ad-hoc situation where I had to adjust the theory based on the task frameworks. If I made the tasks, I could to a larger degree control what they aimed to measure, and the discussion could perhaps yield more conclusive results. Another limitation of not using my own task set, is that I had limited control over the sampling process. Luckily, UDIR was helpful in making sure the sampling was valid for my purpose, and it would not have been possible for me to amass such an amount of data in the short timeframe of my thesis. As with most quantitative studies, the precision and execution of the sampling process is crucial for having reliable data to analyze, and even though I took measures to minimize those limitations I cannot fully dismiss possible effect on the analyses.

My position as a researcher influenced the focus of my thesis, and my consideration of the tasks. Initially I wanted to analyze the tasks in the reading comprehension test as well, but I quickly realized that it would be way too comprehensive. In a study with a larger scope and a larger time frame, investigating the relationships between specific reading comprehension tasks and numeracy might also yield interesting results. Also, when interpreting the tasks, I assumed that all the numeracy tasks should be considered as word problems. This was founded theoretically, but some of the tasks were close to the threshold of being intra-mathematical tasks. Therefore, it might be too generalizing to not identify nuances between the numeracy tasks.

Moreover, I chose some arbitrary intervals of the correlation coefficients. I acknowledge that researchers from other fields might consider the intervals to all be mildly or not correlating, and I considered it a basis for discussion between the tasks. If I chose an artificially high threshold for moderate correlations, I would not be able to present the nuances. Also, reading comprehension is not the primary ability measured by the numeracy part of the national tests, so it was natural to receive coefficients in the range I did. Adding to this, because of the lower number of students in the reading comprehension sub-groups I chose Kendall’s Tau-b over Spearman’s Rho. Therefore, albeit theoretically grounded and reflected upon, some critique can be raised to my choice of coefficient intervals and correlation model.

In the regression analyses it became clear that the models would benefit from adding more predictors, because in general there were only small increases in accuracy when using reading comprehension as a predictor. Especially the issue with low specificity and sensitivity on low and high complexity tasks respectively, which was addressed in-depth, indicated that the predictor models did not fully capture how reading comprehension affects the probabilities of solving the tasks. It did, however, become clear that there was a distinct difference between the ENT and the non-ENT task models when it came to both the explanatory power and the increase in prediction accuracy.

8.2. Didactical implications

There are implications from my research relevant for teachers. I identified some numeracy tasks correlating higher with reading comprehension than others, and a common property of those tasks were the cognitive demand required to solve the task. Therefore, when teachers evaluate student performance, it can be beneficial to assess these tasks outside of the complexity level that they are situated in. Then, teachers can compare student solutions on those tasks to the similar complexity tasks. If a student correctly solves most of the complexity level 3 tasks, but only 1 of the ENT tasks of complexity level 3, it merits some investigation if it is another ability than numeracy which is lacking, such as reading comprehension.

Another implication is how teachers should be aware of the cognitive demand required by the tasks their students are engaging in. It is already discussed in Stein & Smith (1998), but my research supports this notion in the regard that low performers seem to benefit from their reading comprehension ability on tasks with low cognitive demand. By identifying this notion, teachers can apply strategies to expose said students to increasingly more cognitively demanding tasks. It is important that even though the national tests aim to measure the basic skills, my study implies that there are other factors influencing students’ performance. By distinguishing between the mathematical complexity and the cognitive demand of tasks, teachers can be more precise in their assessment of the student performance, and help identifying which areas that should be worked on.

8.3. Own reflections

The process of this thesis has led me down many paths, and what I ended up with did not fully coincide with the original plan I had last autumn. I do believe that I benefitted from being aware of the rigidness in quantitative research from the start, so I made sure every step of the process was well-considered, and in theme with my research question. Also, my personal interest in the subject was crucial to my motivation throughout the process, and I do not think I would have managed as well mentally as I have, had it not been for my personal investment.

The time between determining my research design and getting acceptance from UDIR that I would receive my dataset was specifically challenging. I knew what I wanted to investigate, and how I wanted to approach it, but it relied on an external factor that I had no control over. With my supervisors I developed a back-up plan using a qualitative approach to investigate students thoughts and performance on word problems in a specific classroom setting, but my mind was set on a quantitative approach. Luckily UDIR could provide my desired data, but it was challenging to produce a back-up plan while still hoping for my original plan to pull through.

Looking back at the process, there were things I could have done differently, that maybe would have further enhanced my thesis. In January I spent a lot of time reading relevant theories, but I did not yet have the full picture of how the research was going to be developed. Therefore, a significant amount of that time did not amount to anything concrete, and I ended up broadening my understanding of word problems and reading comprehension. That is not necessarily negative, but in the short timeframe of a master’s thesis the time could have been better spent otherwise. Moreover, I regret the decision of writing the thesis by myself. By discussing with other students, I quickly became aware of the benefits of being able to reflect on the process, construct relevant theoretical background and self-evaluate the steps taken, in a group. Although I did, and do, believe in my own abilities, conducting this research alongside another researcher would undoubtedly benefit both the process and the end product.

Finally, I have had some personal development through this process. I guide students in writing academic papers, but I did not know how hard it would be to implement what I usually guide others to do. Being deep into my own thesis, I struggled through the process with having a metacognitive perspective, and I encountered writing blocks periodically which I battled more or less successfully. However, even though the process was far less smooth than what I would have anticipated, I did end up with a final product that is the culmination of endless hours of work. This leads me to what is maybe the most important part – the ability to trust the process, and that I am able to deliver a product as promised. The process of working over many months towards a set goal line, and then managing to deliver within this timeframe is a crucial skill to have, both in academia and else in the world.

Bibliography

Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education*, *14*(3), 219–234. https://doi.org/10.1207/S15324818AME1403_2

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Bergqvist, E., Theens, F., & Österholm, M. (2018). The role of linguistic features when reading and solving mathematics tasks in different languages. *The Journal of Mathematical Behavior*, *51*, 41–55. https://doi.org/10.1016/j.jmathb.2018.06.009

Blum, W., & Ferri, R. B. (2009). Mathematical Modelling: Can It Be Taught And Learnt? *Journal of Mathematical Modelling and Application*, *1*(1), 45–58.

Blum, W., & Niss, M. (1991). Applied mathematical problem solving, modelling, applications, and links to other subjects—State, trends and issues in mathematics instruction. *Educational Studies in Mathematics*, *22*(1), 37–68. https://doi.org/10.1007/BF00302716

Boaler, J. (1998). Open and Closed Mathematics: Student Experiences and Understandings. *Journal for Research in Mathematics Education*, *29*(1), 41–62. https://doi.org/10.5951/jresematheduc.29.1.0041

Bryman, A. (2012). *Social research methods* (4th ed.). University Press.

Carrell, P. L., & Eisterhold, J. C. (1983). Schema Theory and ESL Reading Pedagogy. *TESOL Quarterly*, *17*(4), 553–573. https://doi.org/10.2307/3586613

Cohen, L., Manion, L., & Morrison, K. (2017). *Research Methods in Education* (6th ed.). Routledge.

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, *19*(4), 497–515. https://doi.org/10.1007/s10260-010-0142-z

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*(4), 405–438. https://doi.org/10.1016/0010-0285(88)90011-4

Flood, J., & Lapp, D. (1990). Reading Comprehension Instruction for At-Risk Students: Research-Based Practices that can make a Difference. *Journal of Reading*, *33*(7), 490–496.

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

Hallett, D., Nunes, T., & Bryant, P. (2010). Individual differences in conceptual and procedural knowledge when learning fractions. *Journal of Educational Psychology*, *102*(2), 395. https://doi.org/10.1037/a0017486

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. - Universitetsbiblioteket i Agder. *American Psychologist*, *58 (1)*, 78–79.

Homan, S., Hewitt, M., & Linder, J. (1994). The Development and Validation of a Formula for Measuring Single-Sentence Test Item Readability. *Journal of Educational Measurement*, *31*(4), 349–358. https://doi.org/10.1111/j.1745-3984.1994.tb00452.x

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding It Up: Helping Children Learn Mathematics*. National Academies Press. https://books.google.no/books?hl=no&lr=&id=pvI7uDPo0-YC&oi=fnd&pg=PA1&dq=adding+it+up+kilpatrick&ots=vC_ab0voak&sig=BvgdP-oiA1CYul7ynNyOzF9jBas&redir_esc=y#v=onepage&q=adding%20it%20up%20kilpatrick&f=false

Kintsch, W. (1988). *The role of knowledge in discourse comprehension: A construction-integration model*. https://psycnet.apa.org/fulltext/1988-28529-001.html

Maagerø, E., & Skjelbred, D. (2010). *De mangfoldige realfagstekstene: Om lesing og skriving i matematikk og naturfag*. Fagbokforl. https://www.nb.no/search?q=oaiid:"oai:nb.bibsys.no:990902659124702202"&mediatype=bøker

Martiniello, M. (2008). Language and the Performance of English-Language Learners in Math Word Problems. *Harvard Educational Review*, *78*(2), 333–368. https://doi.org/10.17763/haer.78.2.70783570r1111t32

Nortvedt, G. (2009). Understanding and solving multistep arithmetic word problems. *Nordic Studies in Mathematics Education*, *15*(3), 23–50.

Nortvedt, G. A. (2011). Coping strategies applied to comprehend multistep arithmetic word problems by students with above-average numeracy skills and below-average reading skills.

The Journal of Mathematical Behavior, 30(3), 255–269. https://doi.org/10.1016/j.jmathb.2011.04.003

Österholm, M., & Bergqvist, E. (2012). Methodological issues when studying the relationship between reading and solving mathematical tasks. *Nordisk Matematikkdidaktikk*, *17*(1), 5–30.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, *96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Perfetti, C., & Stafura, J. (2014). Word Knowledge in a Theory of Reading Comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Rittle-Johnson, B. (2017). Developing Mathematics Knowledge. *Child Development Perspectives, 11*(3), 184–190. https://doi.org/10.1111/cdep.12229

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Schukajlow, S., Leiss, D., Pekrun, R., Blum, W., Müller, M., & Messner, R. (2012). Teaching methods for modelling problems and students’ task-specific enjoyment, value, interest and self-efficacy expectations. *Educational Studies in Mathematics*, *79*(2), 215–237. https://doi.org/10.1007/s10649-011-9341-2

Sfard, A. (1998). On Two Metaphors for Learning and the Dangers of Choosing Just One. *Educational Researcher*, *27*(2), 4–13. https://doi.org/10.3102/0013189X027002004

Statistisk sentralbyrå. (2020). *11980: Elever i grunnskolen, etter region, eierforhold, statistikkvariabel, år og årstrinn. Statistikkbanken*. https://www.ssb.no/statbank/table/11980/tableViewLayout1/

Stein, M. K., & Smith, M. S. (1998). Mathematical Tasks as a Framework for Reflection: From Research to Practice. *Mathematics Teaching in the Middle School, 3*(4), 268–275. https://doi.org/10.5951/MTMS.3.4.0268

Utdanningsdirektoratet. (n.d.-a). *Bokmål – regning, 8. Og 9. Trinn*. Eksempeloppgaver og tidligere nasjonale prøver. Retrieved April 28, 2021, from https://www.udir.no/eksamen-og-prover/prover/eksempeloppgaver-tidligere-nasjonale-prover/8-9-trinn/regning/bokmal/

Utdanningsdirektoratet. (n.d.-b). *Hva er grunnleggende ferdigheter?* Retrieved January 19, 2021, from https://www.udir.no/laring-og-trivsel/lareplanverket/stotte/hva-er-grunnleggende-ferdigheter/

Utdanningsdirektoratet. (n.d.-c). *Kva er nasjonale prøver?* Retrieved April 8, 2021, from https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/om-nasjonale-prover/

Utdanningsdirektoratet. (2017). *Hva måler nasjonal prøve i regning?* https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/mestringsbeskrivelser-og-hva-provene-maler/hva-maler-nasjonal-prove-i-regning/

Utdanningsdirektoratet. (2018a). *Administrere nasjonale prøver*. https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/administrere-nasjonale-prover2/

Utdanningsdirektoratet. (2018b). *Metodegrunnlag for nasjonale prøver*. www.udir.no/globalassets/filer/vurdering/nasjonaleprover/metodegrunnlag-for-nasjonale-prover-august-2018.pdf

Van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Pr.

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*. Swets & Zeitlinger Publishers. https://bibsys-almaprimo.hosted.exlibrisgroup.com/primo-explore/fulldisplay/BIBSYS_ILS71514467380002201/UBA

Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J.-E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology, 28*(4), 409–426. https://doi.org/10.1080/01443410701708228

Walker, D. (2003). JMASM9: Converting Kendall's Tau For Correlational Or Meta-Analytic Analyses. *Journal of Modern Applied Statistical Methods, 2*(2). https://doi.org/10.22237/jmasm/1067646360