Validation of two tests of silent reading comprehension efficiency - Sentence Verification and Picture Selection

Zuzanna Solska

Special Needs Education

40 credits

Department of Special Needs Education

Faculty of Educational Sciences, University of Oslo

Spring, 2023

Validation of two tests of silent reading comprehension efficiency - Sentence Verification and Picture Selection

2023

Validation of two tests of silent reading comprehension efficiency – Sentence Verification and Picture Selection

Zuzanna Solska

Abstract

Background and rationale

The need for early and effective identification of pupils that are at risk of developing reading difficulties has been confirmed both by research (Stanovich, 1986) and the society's expectation formulated by the Ministry of Education and Research in the document *Lærelyst – tidlig innsats og kvalitet i skolen* (Kunskapsdepartamentet, 2017). However, very few reliable assessment instruments are currently used in Norwegian schools (Arnesen et al., 2017). Internationally, one of the most popular methods to identify poor readers is the administration of oral reading fluency tests. The emphasis on this assessment is motivated by the evidence for strong correlations between oral reading fluency and comprehension tests (Hierbert et al., 2012; Kim et al., 2015). Moreover, the administration of these instruments takes a relatively short time and gives the possibility to observe children´s reading behavior. In the last years, however, there was cast doubt about the usefulness of this measure in identifying older children with late-emerging reading difficulties because of early stabilization of oral reading fluency in the development of reading skills (O´Brien et al., 2014). Moreover, individual administration, which is necessary in the case of oral reading fluency measures, takes more resources than a group assessment. The results of the validation study by Johnson and colleagues indicated that the Test of Silent Reading Efficiency and Comprehension (TOSREC) may be an effective and reliable instrument for identifying older children with reading difficulties (Johnson et al., 2011).

The BetterReading research group has noticed the need of introducing assessment instruments that combine the components of comprehension and silent reading rate to the Norwegian context. It led to the development of Picture Selection (PS) and Sentence Verification (SV), which intend to measure silent reading comprehension efficiency – a skill of effective extracting textual information (Simonsen et al., 2022). That inspired the research question which informs this thesis:

Validation of two tests of silent reading comprehension efficiency –
Sentence Verification and Picture Selection

Method

The present study is a part of the BetterReading research project from the Department of Special Needs Education at the University of Oslo, and it concerns the examination of reliability, concurrent validity, and divergent validity of Picture Selection and Sentence Verification. The tests have already undergone pilot testing conducted by Simonsen and colleagues (2022). The BetterReading project has provided data from the children that participate in an intervention program that is a part of the project. Additionally, the author of the thesis and a fellow master student have recruited and assessed children that did not take part in the intervention. The data from 105 children were analyzed for examination of concurrent validity, while scores of 42 children were included for the check of divergent validity and reliability.

Analyses

The bivariate correlations were conducted in the statistical analysis program Jamovi, version 2.3.21 (The Jamovi Project, 2022).

Results and Conclusion

The results indicated that SV has rather low but acceptable for screeners stability of $r = .78$, $p < .001$. In contrast the stability of PS ($\rho = .41$, p < .001) is too low to make this test useful as a screener. Both tests show similar patterns of correlations with other measures of reading skills which may indicate that they assess the same construct that incorporates components of silent reading rate and reading comprehension. The lack of statistically significant correlations between the examined tests and measures of vocabulary, grammar, and non-verbal intelligence suggests that PS and SV have successfully passed the criterion of divergent validity.

Both tests show good construct validity (confirmed by the examination of concurrent validity and divergent validity) but only SV has shown the potential to become a reliable screener in the future.

Preface

There are many who have contributed to this thesis with their time, support, and enthusiasm for the topic.

First and foremost, I would like to express my sincere gratitude to my supervisor, Athanassios Protopapas, who helped me to find a topic for the thesis that perfectly suits my interests and gave me the opportunity to participate in an inspiring research project. Thank you for your generosity with your time and for sharing your knowledge. I also owe acknowledgment to my second supervisor, Kristin Simonsen, for her support and encouragement. I would like to thank also all the members of the BetterReading research group for all the inspiring discussions and practical advice. I deeply appreciate the opportunity to experience how a research team works. Special gratitude goes to my fellow master student, Cecilie Stang, who helped me to collect the data and shared with me all the ups and downs during the writing process.

I am immensely grateful to the teachers that have helped me to recruit the young participants, and to the children that take part in the project. Their time and engagement contributed not only to the data used in this study but also allowed me to get a better understanding of the importance of close cooperation between schools and research field.

Above all, I would like to thank my husband, Valerio D´Epifanio, and my family for their continued support during the five years of my studies and for their unwavering faith in my success. I am especially grateful to my mother, Barbara Solska, who came all the way from Poland to help me take care of my children every time I was writing an exam. I feel very lucky to have so loving people around me.

Thank you!

Oslo, June 2023

Zuzanna Solska

Innholdsfortegnelse

1.INTRODUCTION 12

1.1BACKGROUND AND THE PURPOSE OF THE STUDY 12
1.1.1 MAIN RESEARCH QUESTION: 15
1.1.2 SPECIFIC QUESTIONS THAT ADDRESSED IN THE THESIS: 15
1.2 DELIMITATIONS 16
1.3 STRUCTURE OF THE THESIS 16

2. THEORETICAL AND EMPIRICAL BACKGROUND: READING COMPREHENSION ... 18

2.1 READING COMPREHENSION AS A PRODUCT 18
2.2 COMPREHENSION AS A PROCESS 19
2.3 ASSESSMENT OF READING COMPREHENSION 21

3. THEORETICAL AND EMPIRICAL BACKGROUND: READING FLUENCY 24

3.1 THE CONSTRUCT OF READING FLUENCY 24
3.2 THE COMPONENTS OF FLUENCY IN ORAL MODE 24
3.3 DEVELOPMENTAL PRECURSORS OF FLUENCY IN ORAL MODE 25
3.4 THE MEASURES OF FLUENCY DURING ORAL READING 26
3.5 FLUENCY IN SILENT READING MODE 28
3.6 READING IN DIFFERENT MODALITIES 29
3.9 ASSESSMENT INSTRUMENTS OF SILENT READING RATE 30
3.10 SILENT READING COMPREHENSION EFFICIENCY 36

4. THE METHOD 38

4.1 DESIGN AND DATA ANALYSIS 38
4.2 SAMPLE AND DATA COLLECTION 38
4.3 MEASURES 39
4.3.1 MEASURES USED IN BOTH GROUPS 39
4.3.1.1 SENTENCE VERIFICATION (SV) 39
4.3.1.2 PICTURE SELECTION (PS) 41
4.3.1.4 NEALE ANALYSIS OF READING ABILITY (NARA) 43
4.3.1.5 TEST OF WORD READING EFFICIENCY (TOWRE) 44
4.3.1.6 DISCRETE WORDS 45
4.3.1 MEASURES USED ONLY WITH THE INTERVENTION GROUP 45
4.3.2.1 THE NORWEGIAN VERSION OF THE BRITISH VOCABULARY SCALE (BPVS II) 45
4.3.2.2 TEST OF RECEPTION OF GRAMMAR (TROG) 46
4.3.2.3 RAVEN 47
4.4 RELIABILITY 47
4.5 CONSTRUCT VALIDITY 48
4.6 STATISTICAL VALIDITY 49
4.7 EXTERNAL VALIDITY 49
4.8 ETHICAL CONSIDERATIONS 50

5. RESULTS 51

5.1 DESCRIPTIVE STATISTICS OF VARIABLES 51
5.1.1. SENTENCE VERIFICATION (BOTH GROUPS COMBINED) 53
5.1.2 PICTURE SELECTION (BOTH GROUPS COMBINED) 54
5.1.3 NARA (BOTH GROUPS) 54
5.1.4 TOWRE – WORDS (BOTH GROUPS COMBINED) 55
5.1.5 TOWRE – PSEUDOWORDS (BOTH GROUPS COMBINED) 56
5.1.6 ORF (NON-INTERVENTION) 56
5.1.7. DISCRETE WORDS – ACCURACY (BOTH GROUPS COMBINED) 57
5.1.8 DISCRETE WORDS – RATE (BOTH GROUPS COMBINED) 58
5.1.9. RAVEN (INTERVENTION GROUP) 59
5.1.10 TROG (INTERVENTION GROUP) 60
5.1.11 BPVS (INTERVENTION GROUP) 61
5.1.12 SENTENCE VERIFICATION – TIMEPOINT 1 (INTERVENTION GROUP) 62
5.1.13 SENTENCE VERIFICATION – TIMEPOINT 2 (INTERVENTION GROUP) 63
5.1.14 PICTURE SELECTION - TIMEPOINT 1 (INTERVENTION GROUP) 63
5.1.15 PICTURE SELECTION – TIMEPOINT 2 (INTERVENTION GROUP) 64
5.2 BIVARIATE CORRELATIONS 65

6. DISCUSSION 69

7. LIMITATIONS 75

8. CONCLUSION 76

8.1 IMPLICATIONS 76

REFERENCES: 78

APPENDIX 1: NORMAL Q-Q-PLOTS OF THE VARIABLES 88

FIGURE 19: NORMAL Q-Q-PLOT OF SENTENCE VERIFICATION (BOTH GROUPS COMBINED) 88

APPENDIX 2: SCATTERPLOTS 96

2.1 SCATTERPLOTS – BOTH GROUPS COMBINED 96
2.2. SCATTERPLOTS - INTERVENTION GROUP 102

APPENDIX 3: INFORMATION TO PARENTS AND CONSENT FORM 107

List of figures

Figure 1: The example of task in SV... 40
Figure 2: An example of a task in PS -reading paragraph. 42
Figure 3: An example of a task in PS - a choice of a picture... 42
Figure 4: Histogram of Sentence Verification (both groups combined)................................. 53
Figure 5: Histogram of Picture Selection (both groups combined).. 54
Figure 6: Histogram of NARA - both groups combined.. 55
Figure 7: Histogram of TOWRE-words (both groups combined).. 55
Figure 8: Histogram of TOWRE - pseudowords (both groups combined) 56
Figure 9: Histogram of ORF - non-intervention group... 57
Figure 10: Histogram of Discrete Words - accuracy (both groups combined)....................... 58
Figure 11: Histogram of Discrete Words - rate (both groups combined) 58
Figure 12: Histogram of RAVEN - intervention group 59
Figure 13: Histogram of TROG - intervention group 61
Figure 14: Histogram of BPVS - intervention group... 61
Figure 15: Histogram of Sentence Verification – timepoint 1 (intervention group) 62
Figure 16: Histogram of Sentence Verification – timepoint 2 (intervention group) 63
Figure 17: Histogram of Picture Selection - timepoint 1 (intervention group) 64
Figure 18: Histogram of Picture Selection – timepoint 2 (intervention group)....................... 64
Figure 19: Normal Q-Q-Plot of Sentence Verification (both groups combined) 88
Figure 20: Normal Q-Q-Plot of Picture Selection (both groups combined)........................... 88
Figure 21: Normal Q-Q-Plot of NARA (both groups combined).. 89
Figure 22: Normal Q-Q-Plot of TOWRE – Words (both groups combined) 89
Figure 23: Normal Q-Q-Plot of TOWRE – pseudowords (both groups combined) 90
Figure 24: Normal Q-Q-Plot of ORF (both groups combined)... 90
Figure 25: Normal Q-Q-Plot of Discrete Words – Accuracy (both groups combined)........... 91
Figure 26: Normal Q-Q-Plot of Discrete Words – Rate (both groups combined).................. 91
Figure 27: Normal Q-Q-Plot of RAVEN (intervention group) 92
Figure 28: Normal Q-Q-Plot of TROG (intervention group)... 92
Figure 29: Normal Q-Q-Plot of BPVS (intervention group).. 93
Figure 30: Normal Q-Q-Plot of Sentence Verification – timepoint 1 (intervention group).... 93
Figure 31: Normal Q-Q-Plot of Sentence Verification – timepoint 2 (intervention group).... 94
Figure 32: Normal Q-Q-Plot of Picture Selection - timepoint 1 (intervention group)............ 94
Figure 33: Normal Q-Q-Plot of Picture Selection - timepoint 2 (intervention group)............ 95
Figure 34: Scatterplot of Sentence Verification & Picture Selection 96
Figure 35: Scatterplot of Sentence Verification & NARA 96
Figure 36: Scatterplot of Sentence Verification & TOWRE - words..................................... 97
Figure 37: Scatterplot of Sentence Verification & TOWRE – pseudowords.......................... 97
Figure 38: Scatterplot of Sentence Verification & ORF... 98
Figure 39: Scatterplot of Sentence Verification & Discrete Words-accuracy 98
Figure 40: Scatterplot of Sentence Verification & Discrete Words-rate 99
Figure 41: Scatterplot of Picture Selection & NARA.. 99
Figure 42: Scatterplot of Picture Selection & TOWRE - words... 100
Figure 43: Scatterplot of Picture Selection & TOWRE – pseudowords 100
Figure 44: Scatterplot of Picture Selection & ORF... 101

Figure 45: Scatterplot of Picture Selection & Discrete Words-accuracy 101
Figure 46: Scatterplot of Picture Selection & Discrete Words-rate 102
Figure 47: Scatterplot of Sentence Verification (timepoint 1) & RAVEN 102
Figure 48: Scatterplot of Sentence Verification (timepoint 1) & TROG...............................103
Figure 49: Scatterplot of Sentence Verification (timepoint 1) & BPVS 103
Figure 50: Scatterplot of Picture Selection (Timepoint 1) & RAVEN.................................104
Figure 51: Scatterplot of Picture Selection (timepoint 1) & TROG.....................................104
Figure 52: Scatterplot of Picture Selection (timepoint 1) & BPVS 105
Figure 53: Scatterplot of Sentence Verification (timepoint 1) & Sentence Verification (timepoint 2) 105
Figure 54: Scatterplot of Picture Selection (timepoint 1) & Picture Selection (timepoint 2)106

List of Tables:

Table 1: Descriptive statistics of variables – 1 52
Table 2: Descriptive statistics of variables - 2: Internal consistency.......................................53
Table 3: Pearson’s r and characterization of correlation´s strength..65
Table 4: Bivariate correlations (intervention & non-intervention group)...............................67
Table 5: Bivariate correlations (intervention group): divergent validity 68
Table 6: Bivariate correlations (intervention group): test-retest reliability 68

Word count: 21.128

1.Introduction

1.1Background and the purpose of the study

Reading fluency is traditionally defined as the ability to read accurately, quickly, effortlessly, and with good expression (Armbruster et al., 2001). Assessing fluency based on that definition includes evaluation of accuracy, rate, and often prosody, but it does not involve measures of understanding of a written text. Accurate readers can correctly recognize both familiar and unfamiliar words, by using their knowledge about correspondence between letters and language sounds as well as words that they know “by sight”. An appropriate reading rate should be adjusted to the difficulty of the text and in this way support comprehension (Spichtig et al. 2022). Reading with good speed requires that most of the words are recognized automatically, which enables readers to delegate more of their cognitive resources to understanding the text. Finally, prosody demands attention to punctuation, which is signaled with pauses and raising or lowering the voice. Additionally, good prosody involves to some extent understanding the text because it requires that readers stress words that are central to the text´s meaning and express feelings that are embedded in the content (Mather & Wendling, 2012).

Fluency measures are most often obtained during oral reading because this method is considered to be more reliable in assessing fluency than silent reading tasks due to the possibility of observing reading behavior (Jenkins et al., 2003). What is more, it allows assessing of prosody which can give some impressions of the level of understanding. However, awareness of the difficulty and subjectivity in evaluating prosody often leads to prioritizing speed and accuracy in assessment and instruction. That is unfortunate because fast reading should not be the goal of reading instruction, but rather understanding that is facilitated by the ability to adjust reading speed to the demands of a text (Kuhn et al., 2010). This way of assessing fluency can also result in overlooking problems that are not connected to technical reading. That is especially relevant for later stages of reading development when the correlation between oral reading rates and comprehension decreases (Valencia, Smith, Reece, Wixson & Newman, 2010).

Another drawback of oral reading tasks in school practice is that they need to be administered individually. It makes them time-consuming which restricts the possibility of screening all pupils and constrains time dedicated to instruction. The screening process is meant to identify children that struggle with acquiring reading skills or may be at risk of developing reading difficulties. It helps to provide early interventions tailored to children's needs to remediate difficulties and to prevent the accumulation of problems. A good screener should measure skills that a relevant for the development of reading skills in a given grade. For example, assessing letter knowledge in fifth grade would not be useful in the identification of children at risk for poor performance on the Norwegian national compulsory test (Nasjonale prøver). As every other test, screeners need to have the appropriate level of reliability in order to be useful assessment instruments. However, the purpose of a screener does not require as high reliability as diagnostic tests, and usually, a reliability coefficient of .70 is considered acceptable (Murphy & Davidshoffer, 1994).

After examining the reliability of a future screener, it is important to check criterion validity and classification accuracy. Criterion validity can be determined by the evaluation of correlations between a new screener and established measures of reading skills. The two tests can be administered at the same time point, which gives insight into concurrent validity. However, if the potential screener is administered a longer time before a criterion measure (a reference test) the correlations may give information about predictive validity (Jenkins et al., 2007). The ideal screener would predict with 100% accuracy which children will demonstrate poor performance on a reference test, and which will attain satisfactory scores. However, according to Test Theory, each score on the test is a combination of a "true score" reflecting a person´s skills and a "measurement error", which is a combination of random factors not connected to the targeted skill (John & Martinez, 2014). The test constructors have to compromise between two parameters – sensitivity and specificity. Sensitivity refers to the proportion of children who attained low score on the reference measure and were identified by the screener as being at risk of developing reading difficulties (true positives). Specificity indicates how many children were declared as not being at risk after administration of screener, and in fact, their reading skills are not impaired (true negatives). Unfortunately, the high sensitivity increases the probability of false positives, which results in that the test labels more children as being at risk. It leads to delegation of resources for interventions there where it is not necessary. On the other hand, high specificity increases number of false negatives, which means that children that are in reality at risk of bad performance will not receive the help that they need. (Jenkins et al., 2007).

The importance of early, targeted help for students who fall behind in acquiring academic skills is also emphasized by the Ministry of Education and Research (2017). Although the principle of early intervention is especially relevant for students in grades from 1 to 4, the document underlines that all pupils who fall behind should get adequate help (Kunnskapsdepartementet, 2017). Norwegian national testing program includes compulsory assessment of reading ability with the help of two instruments – Nasjonale prøver (administrated in Grades 5, 8, and 9) and Kartleggingsprøve (carried out in Grades 1-3). The first one is supposed to help with the evaluation of reading instruction and provide the information about reading ability of children of all performance levels. In contrast, the second one gives little information about the reading skills of average or good readers but its´ purpose is to identify pupils that are at risk developing of reading difficulties (Walgermo et al., 2018).

A literature review has shown, however, that there are very few good quality screening instruments that teachers can use to quick assess reading skills of Norwegian pupils more than once per year (Arnesen et al., 2018). One of the few exceptions is the Norwegian adaptation of the Oral Reading Fluency (ORF) measure of the Dynamic Indicators of Basic Early Literacy Skills, which was validated by Arnesen and colleagues (Arnesen et al., 2017). Another Norwegian screener, which has good psychometric parameters is *Ordkjedetest* (Arnesen et al.,2018). Although administration and scoring of ORF takes about 5 minutes per pupil (Johnson et al., 2021), in the bigger scale of a whole school the need of individual administration makes ORF more time-consuming and expensive than group administrated *Ordkjedetest*. The teacher, however, still needs to score *Ordkjedetest* manually for each pupil. According to recommendations of National Center on Response to Intervention (2010) a school should carry out screening process three times per year, which means that individual assessment of all the students and manual scoring takes a lot of time from the instruction. Additionally, the literature review by Arnesen and colleagues (2017) indicates that there is a lack of a reliable and time-saving screeners that could assess silent reading fluency of connected text in the Norwegian context.

The reason for that may be challenges connected to constructing a valid test which is supposed to assess a skill that is not directly observable. While under oral reading fluency assessment the teacher can hear which words are skipped or misread, this is not possible during silent reading. Facing difficulty with assessing elements that are embedded in silent reading fluency, the BetterReading research group at the Department of Special Needs Education at the University of Oslo has proposed a new construct that relates silent reading rate to comprehension of accessible text. Silent Reading Comprehension Efficiency (SRCE) regards the individual optimal reading rate with which textual information from the text of an age-appropriate level can be understood and applied in a simple, practical task (Simonsen et al., 2022). The researchers believe that this skill is important for success in education, a future career as well as for reading pleasure.

This thesis is written in association with the BetterReading research group, and it attempts to validate two assessment instruments that were developed for the project dedicated to achieving a better understanding of SRCE. *Sentence verification (SE)* is an adaptation of the Test of Silent Reading Efficiency and Comprehension (TOSREC), which showed evidence of reliability and validity (Johnson et al., 2021). The children´s task is to decide if single, unrelated sentences are true or false. Like TOSREC, SV can be administered in groups, but our test has the advantage of being implemented as an iPad app that makes scoring even easier. In the second test, *Picture Selection (PS)*, children read one short paragraph at a time and choose the one picture among four presented on the screen that best captures the content. All paragraphs form a coherent narrative. PS as well as SV are time-limited tests that assess children´s performance based on the number of correct responses within a given amount of time.

Validation of these two instruments can reveal their potential for use in Norwegian schools as screeners and may justify their use in further research on SRCE.

1.1.1 Main research question:

Validation of two tests of silent reading comprehension efficiency – *Sentence Verification* and *Picture Selection*

1.1.2 Specific questions that addressed in the thesis:

- What is the concurrent validity of *Picture Selection* and *Sentence Verification* with reliable and valid instruments that are used to assess different aspects of children’s reading skills?
- Are there meaningful differences in the patterns of the correlations of *Picture Selection* and *Sentence Verification* with other tests of reading skill?

- Do the new tests correlate more strongly with measures of reading comprehension or with measures of oral reading fluency?
- What is the stability (test-retest) of *Picture Selection* and *Sentence Verification*?
- Does performance in Picture *Selection* and *Sentence Verification* correlate more strongly with measures of reading skills than with tests of more general abilities (such as nonverbal intelligence and abstract thinking)?

1.2 Delimitations

Creating a reliable and valid test is a complex and multifaced process. The present study is just a humble step in the validation of PS and SV that aims to examine the potential of these instruments for use in research and screening practice in Norwegian schools. Because a valid instrument has to be reliable, the test-retest analysis will be conducted. As a part of criterion validation, the thesis will examine the strength of correlations between the new tests and carefully chosen criterion measures – reliable instruments used in research and diagnostic practice. Because the study does not have a longitudinal design, only concurrent validity will be examined, leaving the check of predictive validity for the future. Investigating the classification accuracy of this test is also outside of this thesis, but in case of positive results of criterion validity examination, it may be considered as the next step in the development of these measures.

1.3 Structure of the Thesis

The first chapter of this thesis has presented the theoretical background and the purpose of the present study. The introduction is followed by research questions that this thesis attempts to find answers to.

The second chapter briefly presents the complex construct of reading comprehension in the light of The Simple View of Reading. Thereafter the chapter presents examples of different ways of assessing the understanding of a written text.

The third chapter starts with a short introduction of the construct of reading fluency and continues with outlining of its components in oral mode. Subsequently, a description of two developmental precursors of oral reading fluency is followed by the presentation of oral reading fluency measures. Later the chapter moves on to the construct of silent reading fluency underlying the differences between reading processes in two modalities. The chapter ends with a description of various methods used in the assessment of silent reading rate, and the introduction of the silent reading comprehension efficiency concept.

The fourth chapter gives insight into the design of the study. It presents the sample and the process of data collection. Subsequently, the chapter gives a detailed description of the assessment instruments used in the study. The chapter ends with a discussion of reliability and relevant forms of validity, as well as ethical principles that were taken into consideration.

The fifth chapter presents descriptive statistics of variables and the results of the statistical analyses.

Chapter six discusses findings in light of the theoretical background, the purpose of the study, and reliability and validity.

Chapter seventh presents the conclusion and outlines the limitations of the present study.

2. Theoretical and Empirical Background: Reading Comprehension

Reading comprehension can be perceived as the ultimate goal of literacy instruction that enables children to acquire new knowledge and participate in modern society (Hierbet & Daniel, 2018). It may be understood both as a product and a process (van den Broek, 2012). The result of the successful understanding of a text is an appropriate *situation model* (also called a *mental model*) that is a representation of the general meaning of a text. Different readers can make various versions of the situation model of the same text depending on their background knowledge and skills. However, a good situation model should always include the central elements of the text and the relations between them (Schwanenflugel & Knapp, 2016).

2.1 Reading Comprehension as a Product

According to the Simple View of Reading differences in the outcome of the process of understanding the text can be explained by variations in two broad skills: decoding and language (or linguistic) comprehension. Gough and Tunmer define skilled decoding as quick, accurate, silent, and context-free word recognition that can develop thanks to the knowledge of letter-sound correspondence rules (Gough & Tunmer, 1986). The second element of the model, language comprehension, allows to access the meaning of the words, sentences, and paragraphs, and integrates them enabling the reader to interpret the linguistic information (Cain, 2010). Lower language skills – vocabulary and grammar – emerge early in development and are learned in great part without conscious effort. They also support the development of higher-level language skills: inferencing, comprehension monitoring, and knowledge about text structure (Hogan et al., 2011).

A reader needs to both decode and have the necessary language comprehension to be able to understand a written text. Empirical studies have supported the theory of the Simple View of Reading providing evidence that both linguistic comprehension and decoding contribute to reading comprehension (Hoover, W. A., & Gough,1990; Protopapas et al., 2012). However, the balance between the contribution of these skills changes during reading development. While decoding has the biggest importance for reading comprehension of younger children, the role of linguistic comprehension increases later (Vellutino et al., 2007; Aaron et al., 1999). Mastering decoding releases cognitive resources that can be delegated to drawing inferences and integrating information from the text with general knowledge (Cain, 2010). This developmental shift is also associated with different expectations that pupils have to meet in the first grades, and later in their educational path. While at the beginning of the literacy instruction, the goal for the pupils is to read words accurately and answer comprehension questions that demand only very simple inferences, the fifth graders are expected to be able to find information in the text and infer different types of relations between text elements. The difficulty of vocabulary and sentence structure in the texts from the curriculum also increases, and the children are expected to use reading as a tool to gain more knowledge (Snow & Vaughn, 2018).

The Simple View of Reading focuses more on skills that readers need to possess to be able to understand the text but does not differentiate between demands that various types of text put on linguistic comprehension and other individual factors such as motivation, stamina, fluency, working memory (Cain, 2010). The responsible assessment of children´s reading proficiency requires however recognition of the fact that different kinds of texts and reading tasks require different skills and levels of text processing.

2.2 Comprehension as a Process

The Construction Integration Model by Kintsch (1988) shows the complexity of comprehension by explaining multiple, cyclic processes during the development of the final representation of the text meaning. Readers need to first decode words from the text, access their meanings and recognize their role in the sentence (phrasing). The first analysis of words' meanings and the structure of the sentences results in temporary, surface representations. Because of limitations of readers´ working memory, meanings of words and phrases are cyclic negotiated with the rest of the text. The meanings of the words that do not suit the context are deactivated, whereas others are strengthened. At the same time, each new word or sentence automatically activates a network of related concepts that may be helpful in interpreting the content of the text. In that process, readers form propositions, which are the smallest units of meaning (Cain, 2010). The network of propositions forms the microstructure of the text. Cohesive devices in the form of anaphors and intercausal connectives help readers to draw necessary inferences about relations between propositions that make microstructure coherent (Kintsch & Rawson, 2005). While anaphors do not contribute with new meaning, but simply refer to a previously mentioned element in the text, intercausal connectives gives additional information to the readers about the nature of relationships between propositions. They may signalize casual (therefore, because), temporal (after, prior to, later), spatial (in front of, back), or contrastive relationship (however, but) (Cain, 2010).

Under reading a connected text, readers also discover relations between different paragraphs and longer parts of the text, which are called macrostructure. The literal meaning of the text, whichcontains basic ideas and connections between them reflected by micro and macrostructure, form the textbase. Creating the textbase may allow readers to reproduce the text but is not enough for a deeper understanding of the text. To create a reach and coherent mental model of a text meaning the readers have to integrate information from the text with their prior knowledge (Kintsch & Rawson, 2005). The final product of reading comprehension processes is influenced not only by the capacity of readers´ working memory and what they know about the topic, but also by their feelings and goals connected to reading activity (Cain, 2010).

The different types of texts and reading tasks might influence children´s goals and determine which processes will be needed to create the appropriate mental model. Constructing the proposition and forming a network between them is required both during the reading of unrelated sentences and whole passages. Children need to recognize words, access their meanings, and apply the knowledge of grammatical structure to understand how the words in each sentence are related (Cain, 2010). Knowing which patterns of words which are allowed in a given language helps to predict the meaning of new words, support fluent reading, and use cohesive devices that help the reader to maintain coherence of the representation of sentence and text meaning (Ecalle et al., 2013). However, comprehension on sentence-level demands only very simple, local inferences about words´ meanings and relations between them. The inferences are made "on-line" (during reading the current sentence) and the network between propositions is rather small, which puts little demand on working memory. The readers do not have to connect multiple anaphors with their references and discover the nature of relationships between many text elements. Therefore, the activation of concepts might be more automatic because readers have still the information from other words or phrases in the sentence available in their working memory (van den Broek, 2012). They might, however, use some simple strategic processes when the content of the sentence conflict with their expectations, and the representation of its' meaning does not meet their standard for coherence, which determines how actively a reader tries to make a text coherent. It can vary not only between people but also may depend on the type of text, situation, and goals of the reader (van den Broek, et al., 2016). However, when children work with longer and more demanding text, many inferences are made off-line, and constraints of the capacity of working memory might limit the access to information needed for necessary inferences (Cain, 2010). In this case, the automatic processes might be not enough to ensured desired standard for coherence, and children need more actively engage themselves in strategic processes, that often need to be learned. The strategic processes might involve rereading the text to find the information that will help to make necessary inferences and using the prior knowledge about the topic or the text structure (van den Broek, 2012). Children have to combine different parts of the text to find its' central elements, make global inferences about the gist of the text, and discover its ´macrostructure (Kintsch & Rawson, 2005).

Various tests of reading comprehension might tap different extent skills and processes that are required for developing of an appropriate situation model. The section below presents some of common test forms used in school and practice research.

2.3 Assessment of Reading Comprehension

In most tests that intend to assess children´s understanding of a text, reading comprehension accuracy is operationalized as a number of correctly answered questions about text content. These tests are usually not timed, and they have increased difficulty level, which helps to evaluate how difficult text children are able to understand. The number of incorrect or missing responses indicates when the text becomes too difficult for children. The instruments have often stop-rules that allow to finish the administration of the test when the child makes predefined number of mistakes. Neale Analysis of Reading Ability (NARA), described in chapter *Measures*, is an example of a comprehension accuracy test.

Much less popular form of reading comprehension tests are instruments that conceptualize comprehension as a rate. These tests may still have increased difficulty of items that challenge readers' understanding but the score is the number of correct responses per given time (Rønberg & Petersen, 2016). An example of a test that attempts to assess both comprehension accuracy and rate is Carlsten. The test is a group administrated maze task. The children need to read silently a coherent text where words are systematically omitted. The pupils are presented with alternatives among which they are supposed to choose a word that suits best to the text

(Carlsten, 2016). The manual of the test does not provide information about psychometric properties of the instrument (Arnesen et al, 2018).

In this section, the focus will be placed on tests that intend to assess comprehension accuracy because of their wide use in school and research practice as well as their relevance for this thesis. Different formats of tests that assess comprehension accuracy vary in which additional skills they tap and how specific answers they require. Additionally, comprehension tests can involve reading in oral or silent mode. Children can also answer the questions verbally, which requires individual administration, or in writing, facilitating group administration.

On the one end of the scale, there is a task called retell, which demands children to read a story and after that tell the test administrator what the text was about. This form of assessment gives a lot of freedom to the children, who can themselves choose which elements and relations to include in retelling. On the other hand, this type of instrument puts high demands on expressive language, narrative skills, and memory. It may be also difficult to determine why some elements from the text do not appear in children´s retelling. Some elements could be misunderstood during text reading, forgotten, or falsely perceived as not important. The children may also have problems finding the right words to reproduce the content of the text (Cao & Kim, 2021).

Another form of assessment of the comprehension is a test with open-ended questions. Different items from such a test usually tap various skills from memory for detail to getting the gist of the text. Therefore, they are not always perfectly intercorrelated – they can measure different things that are included in a comprehension construct. Both open-ended question tests and the retell format are difficult to score and the person administering the test is usually equipped with a detailed guide with examples of correct and incorrect answers.

Cloze tasks and multiple-choice tests target much more specific elements than retell and do not put as big demands on memory as two other tests. They are also easier to score and can be easily applied in a group setting. Cloze tasks are constructed by systematically removing every 5th or 10th word from the text. Children are provided with three or four alternatives of words, and they need to choose the one that best suits the sentence. The items can tap different skills, like grammar, vocabulary, general knowledge, or even spelling. Giving the right answer, however, rarely requires understanding more than one sentence (Cain & Oakhill, 2006). In contrast, a well-constructed multiple-choice test can tap understanding of phrases, sentences, as well as entire passages. The children´s task in this case is to select one of three or four prespecified responses to a question, which minimizes demands on verbal skills and memory (Cain, 2010).

Another method of assessment that does not requires a verbal response and are suitable for group administration is the recognition of true/ false sentences. After reading a passage, children are presented with sentences connected to the content of the text. Children’s task is to judge if the sentences correctly reflect the meaning of the text. This type of test may be useful in assessing children’s ability to remember details, but it gives room for guessing because there are only two alternatives (true sentence/ false sentence). Recognition of sentences that are consistent with the mental representation of the text does not give information about pupils´ ability to make inferences by themselves (Cain & Oakhill, 2006).

3. Theoretical and Empirical Background: Reading Fluency

3.1 The construct of Reading Fluency

Although there is a consensus among researchers that fluency is a necessary element of skilled reading both in oral and silent mode (Kuhn et al, 2010; Hudson et al., 2009), there is no one universal definition of the construct (Schwanenflugel et al, 2016; Mather et al., 2012). In the last two decades perception of fluent reading was strongly influenced by National Reading Panel which describes it as “the ability to process text quickly, accurately and with proper expression” (National Reading Panel, 2000). The elements that are often included in the definition of fluency are *accuracy, automaticity,* and *prosody* (Kuhn et al., 2010; Mather & Wendling, 2012). Some authors also include a component of *comprehension* of connected text in the construct of fluency (Mather & Wendling, 2012).

3.2 The Components of Fluency in Oral Mode

Although the first two ingredients of fluency – accuracy and automaticity – develop in parallel, the second one is less constrained than the first one. It means that children acquire accuracy faster, and later in the development of reading skills there are smaller individual differences in accuracy than in automatic word reading. Moreover, the goal of literacy instruction is to learn to read all the words, even the novel ones, accurately, while automatic reading of a specific word can be achieved only by experience with that written word (Paris, 2004; Schwanenflugel, & Kuhn, 2016). Accuracy, defined as the correct identification of printed words requires letter knowledge and understanding that they can be converted into speech sounds (phonemes) that build a word (Fletcher et al., 2019). In the beginning of literacy instruction, children rely on phonological strategies to read words. They identify letters and connect them to phonemes, then sound them out, which allows them to accurately read regular words and pseudowords (Mather & Wendling, 2012; Price et al., 2016).

According to the self-teaching hypothesis, practicing oral reading, especially with the help of a skilled reader, helps children to direct their attention to letter strings and their phonological representation. This allows them to reinforce knowledge about relations between phonemes and string of letters and acquire information about how to read irregular words (Cain, 2010).

This repetitive experience with printed words allows developing of the second component of fluency, that is, automaticity. The larger chunks of letters, and later all the words undergo unitization and can be read “bysight”. It means that the words are recognized almost instantly and effortlessly, which allows for faster and smother reading (Schwanenflugel & Knapp, 2016). According to Ehri, not only the pronunciation but also the meaning of sight words is activated (Cain, 2010), facilitating comprehension. Moreover, Logan (1997) argues that in addition to speed and effortlessness, the notion of automaticity also implies that reading words is not intentional and does not demand consciousness. That allows for releasing cognitive resources to comprehending the text. Although readers never stop to use the phonological strategy, especially when they encounter unfamiliar words, use of the sight-word mechanism increases with development of reading skills.

The third component of oral reading fluency is *prosody,* also called *prosodic expression* or *reading expression*. Reading with good prosody is characterized by changing of intonation, assigning stress on some words, and pausing, which reflects children´s attention to punctuation and the meaning of the text. Therefore, prosody can give some insights about readers’ comprehension. Pupils can usually read with appropriate use of vocal elements when they have already developed to some extent their accuracy and automaticity. Good prosodic expression can also support storing information from the text in working memory, and in this way facilitate comprehension (Schwanenflugel & Kuhn, 2016).

3.3 Developmental Precursors of Fluency in Oral Mode

Because of the influence of the double-deficit hypothesis, two underlying cognitive skills are most often mentioned in the literature about oral reading fluency, namely phonological awareness and rapid automatized naming (RAN). Both skills are strong predictors of reading performances, and their assessment even before the initiation of reading instruction can help to identify children at risk of developing poor reading skills (Norton & Wolf, 2012). According to the double-deficit hypothesis, impairment of one or both of these skills is a main cause of reading difficulty on the word level (Elliott & Grigorenko, 2014). Phonological awareness refers to a metalinguistic ability that allows children to understand that words are built from language sounds (phoneme), such as syllables and phonemes, and to manipulate them. This skill is a good predictor of reading fluency also after controlling for other factors (Cain, 2010). Phonological awareness is connected to skills that are crucial for young readers: phoneme segmentation, which allows to divide a word into individual language sounds, and phonological decoding, that is, reading words by sounding out each letter and putting all the sounds together (blending) (Price et al., 2016).

Rapid automatized naming is assessed by a task that requires quick and fluent naming of well-known visual stimuli that are presented simultaneously in an array. RAN was often perceived as a skill dependent on phonological awareness because it requires accessing phonological codes from memory. However, it also has a component of time in which the task should be done (Bar-Kochva, 2013). Moreover, the scores on tasks that intend to measure phonological awareness are only moderately correlated with RAN-tasks, and both skills contribute uniquely to reading ability. Additionally, results of neuroimaging studies indicate that there are differences in brain activation during solving RAN-tasks and tasks that intend to measure phonological awareness. All of that suggest that RAN is a construct more independent from phonological awareness that it was believed before, and its' role importance increases after children master accurate decoding of words (Norton & Wolf, 2012). In the past the role of RAN in development of reading skills were also considered to be bigger in nontransparent ortographies, which have inconsistent realations between letters and language sounds (Schwanenflugel & Khun, 2016; Bar-Kochva, 2013). Cross-linguistic studies has shown that RAN may be a predictor of future reading fluency across different types of orthographies (Norton & Wolf, 2012).

3.4 The Measures of Fluency During Oral Reading

Most measures of oral reading fluency used in school practice and research contain only two elements: accuracy and automaticity. Prosody is rarely included in the assessment because it is difficult to assess objectively. Ratings schemes or checklists can vary considerably in the number of dimensions that are evaluated and they do not have enough precision (Mather & Wendling, 2012).

Therefore, most researchers and teachers use two types of individually administered instruments that intend to assess fluency: *word* or *pseudoword list fluency* (called also *word reading efficiency*) and *passage fluency* (*oral reading fluency*). The first kind of test requires that children read orally words or pseudowords from a vertical list as fast and accurately as they can during a prespecified amount time. The administrator marks items that are misread and computes the score by subtracting them from the number of all read words (Schwanenflugel & Knapp, 2016). An example of this instrument is TOWRE, which is described in more detail in the methodology part. In the second type of instrument, children are asked to read aloud connected, grade-level texts for a specific amount of time. Also here the readers are being instructed to read fast and accurately. The score is also the number of correctly read words per unit time (often 1 minute) (Schwanenflugel & Knapp, 2016). In this study, Oral Reading Fluency (ORF), validated by Arnesen and colleagues (Arnesen et al., 2017), was used to examine passage fluency. A more extensive description of the test can be also found in methodology part.

The scores attained on both type of fluency tests are influenced by cognitive and motor skills connected to speech and eye movements during reading (O´Brien et al., 2014; Price et al., 2016; Hierbert et al.; 2012). Therefore, they are not "clean" measures of accuracy and automaticity. However, their high correlation with comprehension lead to accepting fluency measures as indicators of general reading competence and focus on fluency intervention and assessment in schools (Kim, 2010; Denton 2011). Although word fluency lists and passage fluency are not enough to diagnose reading difficulties, they are often used as screeners due to good predictive validity (Mather & Wendling, 2012). They were proven to be helpful in identifying pupils who are at risk of weak performance on compulsory periodic comprehension tests (Denton, 2011). In addition to good accuracy in identifying poor readers, teachers have noticed the practical benefits of these tests. They have relatively short administration and scoring time and can be used repeatedly, which gives the possibility to use them to monitor the progress of the pupils (Wissinger, 2023). Moreover, the person administering the test can observe children's reading behavior and know if they are staying on task (Price, 2012). Although prosody usually is not systematically rated in that instrument, the tester may get some impressions about children´s attention to the punctuation marks, timing, and phrasing. Hintze and colleagues note also that measures of fluency can be perceived as more objective than comprehension measures because they are less influenced by socioeconomic and racial background (Hintze et al., 2002).

However, the evidence of the weakening correlation between oral reading fluency and comprehension in later stages of literacy development brought into question the legitimacy of the dominance of oral reading in assessment and instruction for all age groups (Spichtig et al., 2016; Psyridou et al., 2022). There are two possible explanations for that phenomenon. Firstly, oral reading fluency is a semi-constrained skill and stabilizes earlier than comprehension, which develops through the whole lifespan (O´Brien et al., 2014; Paris, 2004). Secondly, texts from the curriculum increase their difficulty, and according to Chall´s reading stages, after third grade, it is expected from pupils to use written information to learn (Schwanenflugel & Knapp, 2016). It puts bigger demands on the use of higher-level text processing, which involves inferencing, monitoring of understanding, and using knowledge about text structure. Therefore, oral fluency interventions may have a limited effect on the general reading skills of older students (Fletcher et al.,2019). Denton and colleagues have found evidence of significant, positive associations of moderate strength between comprehension and oral reading fluency in 6-8 Grades. The results have, however, confirmed that these correlations are weaker than associations observed in first grades of primary school (Denton, 2011).

Another concern regarding the dominance of oral reading fluency in school practice is the weaker performance of today's pupils on tests that involve both silent reading rate and comprehension, compared to scores attained by children in 1960 (Hierbert & Daniel, 2019). A possible explanation of these results is too big focus on speed during assessment and remediation of fluency, which does not leave enough time for the guidance of the silent reading process and instruction in reading strategies (Hierbert, 2012).

The third argument against the monopoly of oral reading fluency in school practice concerns the relevance of assessment and instruction to requirements that pupils will meet in future education and career. Although oral fluency intervention may impact positively the text understanding of younger pupils, the effectiveness of the intervention is limited when it comes to comprehension of the older children (Wexler, Vaughn, Edmonds, and Reutebuch, 2008). Therefore, this type of intervention may not help young readers in developing effective silent reading skills that are beneficial in modern society.

3.5 Fluency in Silent Reading Mode

According to Khun and colleagues (2010), silent reading fluency, similarly to oral reading fluency, involves accuracy and automaticity that facilitate comprehension of a text. Because vocalization is not the outcome of functional silent reading, prosody is usually perceived as an element only of oral reading fluency (Kuhn et al., 2010). However, in literature, there is a notion of *implicit prosody,* which is an inner voice that children develop during the transition from oral to silent reading. Implicit prosody is supposed to help readers to analyze the structure of the sentence and the relations between words facilitating comprehension (Webman-Shafran, 2018). Some findings also confirm the use of that inner voice among adults (Kuhn et al., 2010).

Although implicit prosody may be an intriguing problem for research, it will probably not be used in the assessment of pupils at school, thus it will not be further discussed in the thesis. Despite the fact that oral and silent reading fluency include the same components – accuracy and automaticity – and their measures correlate strongly with each other (Hierbert & Daniel, 2019), there are some arguments that allow perceiving them as two separate constructs. They concern differences in the input of underlying cognitive skills and the nature of reading process in unlike modalities.

3.6 Reading in Different Modalities

Oral reading with support of an adult may be perceived as a natural transition from listening to reading by a skilled reader. It gives the possibility for interaction and helps younger children to reinforce their knowledge about the relation between sounds and letters, staying on task, and supporting comprehension. However, older readers more often chose silent reading as a more effective tool for extracting information from a text (Price et al., 2016; Price et al., 2012). This is probably because the silent reading rate, in contrast to the oral reading rate, is not slowed down by speech production and the reader can read more words when a text has the right difficulty level (Hierbert et al. 2012). Such a text should be possible to read 99% accurately by the pupils and contain familiar words and topics (Mather & Wendling, 2012). However, children first need to become fluent oral readers to transition to silent reading between fourth and fifth Grade, but even then, they can demonstrate different levels of comprehension in two modalities (Price, 2012; Denton, 2011). According to Chall, after the transition to silent reading mode, fluency is developed enough to release cognitive resources that allow children to learn from the text that they read (Spichtig et al., 2016). That is important also because silent reading puts bigger demands on children´s motivation and ability to monitor text understanding since pronouncing the words can no longer help with storing information in short-term memory. Lack of this support can also influence the time that children can stay on task.

It seems that RAN and phonological awareness are important developmental precursors for reading fluency in oral as well as silent mode. However, they can play somewhat different roles during the acquisition of those skills. Phonological skills are crucial for developing reading skills in general. However, they are more highly activated during oral reading due to the necessity of pronouncing words (van den Boer et al, 2014; Price et al., 2016). Similarly, naming speed is more strongly associated with oral reading (van den Boer et al., 2014). This is because

both oral reading and RAN require articulation and fast retrieval of connections between phonological representations and visual stimuli (van den Boer et al, 2014; Price et al, 2016).
However, RAN is also believed to play a role in developing orographic knowledge (information patterns of letters that represent language sounds in a specific language), which is important both in oral and silent reading. Moreover, there is evidence that this skill is a good predictor of silent reading fluency in the early years of primary school (Bar-Kochta, 2013).
On the other hand, studies by Price and colleagues have shown that although RAN is associated with oral reading fluency in fourth graders, it was not associated with silent reading fluency. This may indicate that RAN does not play an important role in supporting silent reading fluency in later grades of primary school (Price et al. 2016).

However, there is evidence of correlations between oral and silent reading fluency measures that may be explained in two ways. Firstly, both silent and oral reading fluency require automaticity. If children lack that skill under oral reading, it will be also visible in the results of the silent reading test, because ineffective recognition of words will impact cognitive resources needed to process the text content. Secondly, tests that intend to assess fluency in both modalities are timed, therefor poor students will not be able to fulfill either of the assessments (Hierbert & Daniel, 2019). Moreover, there is evidence that oral reading fluency supports the development of silent reading fluency (Price et al.,2016).

Strong correlations between oral and silent fluency measures allow one to hope that both are equally good indicators of general reading skills. Additionally, the lack of requirement of pronouncing words during silent reading assessment makes it possible to administer in a group setting, which on a large scale can save a lot of resources that can be delegated to instruction and remediation of reading difficulties. Therefore, it is worth considering if silent reading fluency measures could be used in the assessment of older pupils.

3.9 Assessment Instruments of Silent Reading Rate

There is a variety of measures that intend to assess *silent reading rate* (or *silent reading fluency*). They differ, however, in terms of emphasis on checking reader´s understanding of the text, observation of reader behavior, the impact of underlying skills on the outcome, methods of administration, the time provided to students, type of task, and variation in item difficulty.

One feature of the instruments that may greatly impact the results of the assessment is whether the tests directly check comprehension. An inexpensive and timesaving method that can be easy to apply in a group setting involves children reading text or a list of words and marking the last read word when the time is up (Price et al., 2012; Denton et al., 2011). The results, however, do not give information about words that have been skipped by the child or read more than once, which leads to questionable accuracy of the measure. Furthermore, children may fail to correctly report the last word that they have read. If the reading task is not followed by comprehension questions, it is also impossible to evaluate readers´ understanding of the text. Spichtig et al. (2022) explain that, in the case of *superficial reading,* pupils read fast but fail to engage in the text because of insufficient skills or weak motivation, which leads to weak comprehension. Good readers maintain quite a stable silent reading rate throughout paragraphs and slow down when they encounter more challenging parts of the text. In contrast, children with reading difficulties tend to read the first parts of a passage at a more or less appropriate rate, but they unnecessarily speed up reading later paragraphs (Hierbert & Daniel, 2019). What is more, the poorest readers can perform "fake reading", which means that they just pretend to read a text (Griffith & Rasinski, 2004).

Therefore, most researchers choose methods of assessing silent reading rate that include a comprehension check. The choice of how to assess comprehension can, however, have great consequences for research outcomes and the evaluation of pupils´ reading skills. Measures of silent reading fluency vary greatly in terms of comprehension units and level of text processing that are taken into consideration.

One group of tests measure rate in relation to recognizing individual words. These instruments put the smallest weight on comprehension and involve a low level of text processing. However, they contain some decision component. For example, in the *word reading fluency task* from the Finnish test battery *ALLU* (*Reading Test for Primary School*) pupils are asked to read silently 80 sets of four phonologically similar words and connect one of the words with a picture that represents it. Children are given 2 minutes to do as many examples as they can and the results reflect both silent reading fluency and accuracy in choosing a correct alternative (Psyridou et al., 2022). Silent word reading fluency tests can also involve a lexical decision task where children need to read through a list of words and pseudowords under limited time. Pseudowords (also called nonwords) are a combination of letters that can be pronounced according to the phonological rules of the language, but they do not bear any meaning (Mather & Wendling, 2012). The children’s task is to go through as many items as possible and simultaneously cross out pseudowords. In this case, the outcome is computed as the number of read words and pseudowords minus the number of errors (van den Boer et al., 2014).

TOSWRF (*Test of Silent Word Reading Fluency*) is an example of the third popular method, called *slasher* or *word chain*, in assessing individual word reading speed in silent mode. This instrument displays strings of unrelated words written in lowercase without spaces. Pupils are supposed to divide strings into individual words by drawing lines between them. The test´s time limit is 3 minutes, and the results an indication of the speed at which children can identify individual words (Denton et al., 2011). A similar instrument, *ordkjedetesten*, is used in Norwegian schools. The description in the manual says that the primary purpose of the test is to assess children’s decoding skills, but since only 4 minutes are provided to separate words in 90 strings, the test can also give an indication about pupils´ silent reading rate (Høien & Tønnesen, 2008).

All these instruments require only paper and pencil, and are inexpensive, easy, and fast to administrate and score. They can also be suitable for a class setting. The instruments differ, however, in the difficulty of items. While ALLU contains words that should be well known to children and the test used by van den Boer et al. (2014) target only short (bisyllabic) words, items from TOSWRF are characterized by increasing difficulty up to adult-level vocabulary.

The second group of instruments that intend to measure silent reading rate use the sentence as a comprehension unit. In this approach, children are supposed to read short, grammatically simple sentences one by one and judge their truthfulness, by marking one of two alternatives: “correct/true” or “incorrect/false”. The test has a time limit, and the number of correctly verified sentences within the time limit is the child’s score. The content of the sentences should be easy to understand for children in a specified age range. An example of such a test administered in paper and pencil format is TOSREC (Test of Silent Reading Efficiency and Comprehension). Similarly to instruments that take the word as a comprehension unit, sentence verification tests are easy to administer in groups and time-saving (Johnson et al., 2011).

Another group of instruments measure the rate of silently reading words that appear in a meaningful context that is larger than a sentence. The Test of Silent Contextual Reading Fluency (TOSCRF) is quite similar to TOSWRF, but the words children have to separate from each other build a coherent text. However, all the letters are written in uppercase and there is no space between words or punctuation that could indicate the structure of sentences. As in the TOSWRF, children have 3 minutes to draw lines between as many words as they can, and the number of correctly separated words constitutes the score (Denton et al., 2011). Although children are presented with coherent text, it is difficult to say how great an understanding of a paragraph or sentence is required to successfully separate words from each other. This test is thus placed somewhere between instruments that use single words, sentences, and paragraphs as a comprehension unit.

Similar uncertainty as to the level of comprehension needed to fulfill the tasks may concern maze tasks. In this type of test, children are supposed to read a coherent text in which words are systematically deleted. Usually, every 7th or 10th word in each paragraph is omitted. Children need to choose the one out of three proposed words that best fits the text (Kim et al., 2015; Denton et al., 2011; Wissinger et al.,2023). Although the correct answers can require an understanding of a bigger part of the text, it is usually enough for a student to grasp the meaning and structural requirements of one or two sentences to choose the right alternative (Schwanenflugel & Knapp, 2016). Children are allowed quite a short time (about 3-4 minutes) to do the task to ensure that no one will be able to read all the text. Since the emphasis is put on the rate, not on the comprehension, the choice between alternatives is made quite easy.

None of the methods of assessing rate during silent reading presented above requires extensive use of higher level language skills (comprehension monitoring, use of text structure, inferencing) or makes great demands on working memory, since all the comprehension tasks are fulfilled on-line (during reading the text). Children do not have to engage in a higher level of text processing that is necessary to build a complex and accurate *mental model* – the representation of the text´s meaning (Hogan et al., 2011). This stands in contrast to the real world demands of reading in which assimilation of information from a written text much longer than one sentence is crucial for functioning in society.

The third group of instruments intends to measure a construct called comprehension-based silent reading rate, which can be described as the interplay between the rate and comprehension of longer, intact, and accessible texts read in silent mode (Hierbert et al.,2010). The children's task involves reading a longer text and answering follow-up comprehension questions that are often in multiple-choice format. The administrator controls how much time children spend on reading the text (Hierbert et al.,2010) or children get limited time to go through the text. They usually do not have the possibility to go back to reread the text. To ensure that assessment captures individual differences where the emphasis is put on rate, not on comprehension, some researchers use minimal comprehension level that is required to use gathered data in studies. Often results are considered valid only if children demonstrate at least 70 % comprehension level (Taylor,1965; Spichtig et al., 2016; Rasinski et al., 2011; Hierbert et al., 2012).

To ensure the rule of minimal acceptable comprehension level, researchers can discard data that do not fulfill that demand, or they can use an adaptive assessment that adjusts the difficulty of items to children's abilities. An example of such an instrument is the web-based instrument InSight, which changes the initial text in response to pupils' vocabulary grade level (Spichtig et al., 2022). Hierbert and colleagues argue, however, that it difficult to establish one universal threshold for comprehension, because the type of a text determines what proportion of accurate comprehension is needed to ensure that reading is efficient (Hierbert et al., 2012).

Tests developed to assess CBSRR are more similar to compulsory assessments in Norwegian schools and to dealing with a text in real life than instruments that use smaller comprehension units. This is because readers do not need to stop reading to do comprehension tasks and they need to use their higher-level language skills to integrate content from different parts of the text with their general knowledge. They can also be applied in the classroom or used in research in combination with other methods that include more advanced technological solutions.

For example, the use of eye-tracking technology, which records eye movements during reading, can give detailed information about reading behavior. That includes data that show how long readers focus on one word (fixation), how often they skip a word, and when they need to go back to a previous part of the text (regression). Readers´ behavior can indicate their understanding of the text but the method is often perceived as unnatural, lacking ecological validity, expensive, and difficult to administer in a group setting. Additionally, children need to be able to carry out the task without moving their head too much, to ensure that the collected eye tracking data are precise.

This latter requirement is not needed when the window method is used, in which children are asked to read a text or a list of words that is gradually exposed as they press a button to move on. The window can reveal different text units – from a word to a paragraph. This approach allows simultaneous assessment of many children but gives less precise information about reading behavior, which limits inferences about children’s comprehension of the text if the reading task is not followed by questions about the text. The method that involves the additional behavior under reading (pressing the button to expose the next segments of the text) may appear unnatural.

Price et al. (2012) point out that another self-paced method—underlying the text under reading—is much more similar to the task children usually do at school. Pupils are supposed to read a passage presented on the tablet and simultaneously underline words from the text with a stylus. If children go back to previous words or sentences, they should underline re-red words once again. This allows the program installed on the tablet to monitor reader´s rate, and record pauses and regressions. The assessment also includes comprehension questions children are supposed to answer after completing the reading task (Price et al., 2012).

When using technology to assess silence reading rate, it can be relevant to take into consideration how reading on-screen affects children´s performance and reliability of assessment. Educational and professional success in the global-digital age requires efficient dealing with texts also in digital format and children in Norway are well used to reading on a tablet or in higher grades on PC. Researchers note also the value of technological solutions in the remediation of reading difficulties in their adaptivity and flexibility that facilitate scaffolding (Hierbert et al., 2010).

Therefore, assessment with the use of technology may be as natural for children as the traditional paper-pencil method. However, the choice of the method can have an impact on the outcomes of the assessment. In the study by Hierbert et al. (2010) fourth-grade pupils read text on screen faster than on paper, but they demonstrated a similar level of comprehension. The results of the systematic literature review conducted by Singer and Alexander indicates, however, that comprehension may be compromised when reading a digital text that is longer than 500 words or one (screen) page (Singer et al., 2017). Results of other study indicates *screen inferiority* that involves shallower learning processes when the text is presented on screen and overconfidence of readers about their performance (Mangen et al., 2013). Lauterman et al. (2014) also point out potential influence of children’s personal preferences about text reading, and of guidance from a professional that can help pupils engage more deeply with reading on screen.

In addition to features as adaptivity, potential accuracy, and authenticity, digital assessments are considered useful in classroom settings because of the possibility of group administration and automatic scoring. However, poor readers can perform much worse in assessment that combines group administration with a digital format compared to one-on-one paper-and-pencil testing under the observation of an examine (Hierbert & Daniel, 2019). This indicates that modality and way of administration influence the reading behavior of poor readers.

The type of test of silent fluency and different ways of administration may determine which underlying skills and factors affect the results.

The ability connected to aforementioned higher level language skills, especially to monitoring of comprehension, is stamina that can also have an impact on the outcome of the assessment. It allows the reader to keep attention, interest, and proficiency at an appropriate level through whole the text. Therefore, stamina is crucial for applying reading strategies and monitoring ongoing understanding (Hierbert et al., 2010). That is why stamina can play a greater role in group-administered tests of independent reading that use longer texts as comprehension levels where children do not receive any help in staying on task. Reading entire texts and answering comprehension questions afterward also puts greater demands on memory than decision tasks that children do immediately after reading a word or short sentence. Some measures, for example, the battery from ALLU, requires also more intensive visual processing than other instruments.

3.10 Silent Reading Comprehension Efficiency

The previous chapter was titled “Measures of silent reading rate” rather than “Measures of silent reading fluency” in order to signal the discrepancy between elements of the theoretical construct of fluency and what the instruments really assess. While it is possible to evaluate the accuracy, reading, reading speed, and, to some extent, prosody during administering *pseudoword list fluency* and *passage fluency,* measures of silent reading rate cannot give a direct picture of these elements. They need to employ a comprehension component to assess how accurately, and fast children read a text silently. Therefore, Hierbert and colleagues have chosen to introduce the notion of comprehension-based silent reading rate (CBSRR), which involves reading with at an appropriate rate that supports comprehension of connected text with use of the higher level language skills (Hierbert et al., 2010). However, measures of CBSRR include not only components of rate and decoding but also the impact of memory. During tasks that are supposed to measure CBSRR children are not allowed to reread the text before answering the questions. This does not seem to resemble a natural reading situation, where one of the most basic strategies for comprehension monitoring is going back to previous parts of a text. Therefore, researchers from BetterReading propose another term, which is silent reading comprehension efficiency. It encompasses the time in which a reader can extract information from an accessible text during silent reading in a natural situation to use it further in a practical task. We believe that effective silent reading, where the rate is moderated by strategies that guard comprehension, is an important skill for education and later career. This thesis aims to examine if the construct is in fact measurable by two tests – Picture Selection and Sentence Verification – and how these tests are associated with other measures of reading skills.

4. The Method

The present study aims to take the first step in the validation of two assessment instruments – *Picture Selection* (PS) and *Sentence Verification* (SV). The instruments were developed for the research project “BetterReading”, and this study is conducted in association with that project.

4.1 Design and Data Analysis

This thesis does not have ambitions to explain possible causal relationships between comprehension, fluency, and silent reading comprehension efficiency, but solely investigates if there are relations between PS, SV, and other measures of reading skills, vocabulary, grammar, and non-verbal intelligence. Therefore, this quantitative study has a non-experimental design and will not involve manipulation of variables (Kleven, 2002). The description of relations between the scores from different measures will be based only on the strength and significance of statistical correlations. The analysis will be conducted with the help of the statistical analysis program Jamovi, version 2.3.21 (The Jamovi Project, 2022).

4.2 Sample and Data Collection

The original sample included 111 children in Grade 5 attending elementary school in the Oslo area. Because of missing data or procedural mistakes, 6 children had to be removed from the sample.

There were two groups (sources) of participants: 47 students received a reading intervention that was a part of the BetterReading project. These children were assessed several times by research assistants connected with the project, while 63 children who do not receive intervention were assessed just once by the author of this thesis and a fellow master student. The students were also engaged in recruiting the participants. All participants had to meet a criterion of attending Norwegian school from the first grade. Additionally, children from the intervention group needed to be able to read 40 words per minute with at least 70% accuracy, secured with preliminary testing using word reading efficiency and oral reading fluency tests. All the children were assessed with the following tests: Sentence Verification (SV), Picture Selection (PS), Oral Reading Fluency (ORF), Discrete Words, Test of Word Reading Efficiency (TOWRE), and Neale Analysis of Reading Ability (NARA). However, the intervention group was tested just with one text from ORF, and that is why their data was not taken into consideration under the analysis of internal consistency. Additionally, the intervention group was assessed with Raven's Color Progressive Matrices (CPM), the Norwegian version of the Test of Reception of Grammar (TROG), and the Norwegian version of the British Vocabulary Scale (BPVS). One individual assessment session lasted about 45–70 minutes per participant and took place in a group room at school. All research assistants and master students working on the project have received training in test administration. The sessions were audio recorded which gave the possibility to check children´s answers if there was any uncertainty connected to the scoring.

4.3 Measures

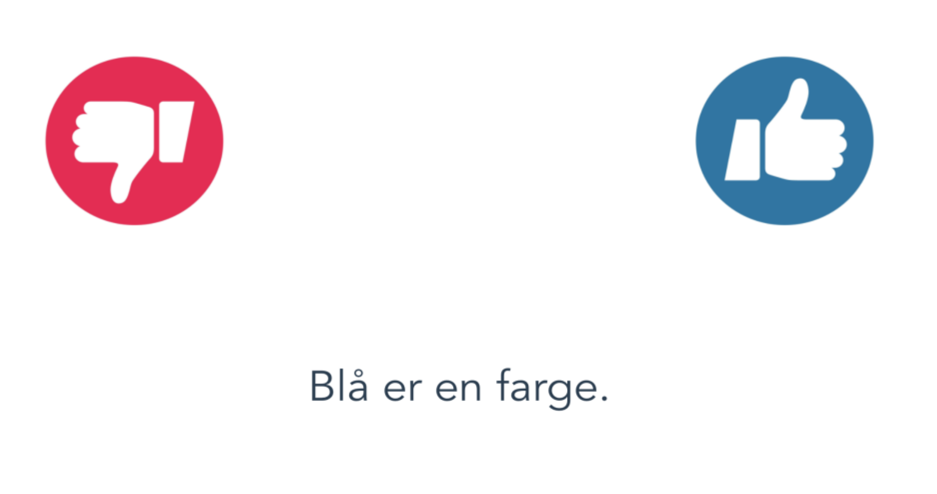
4.3.1 Measures Used in Both Groups

4.3.1.1 Sentence Verification (SV)

Sentence Verification (SV) is a computerized test that was developed for the BetterReading project for grades 2 and 5. This thesis aims to validate the version of SV for the higher grade. The construction of this assessment instrument is inspired by the Test of Silent Reading Efficiency and Comprehension (TOSREC). Validation of the American version of TOSREC has shown potential for the use of this kind of test in the screening process and research (Johnson et al., 2011). These promising results have caught the attention of researchers from the BetterReading project, who aim to develop a similar instrument for the Norwegian context.

The test is implemented as an iPad app. During administration unrelated sentences appear one by one on a tablet. Children's task is to read them silently and judge their truthfulness, by tapping on an icon with either a thump up (true sentence) or an icon with a thumb down (false sentence). An example of a true sentence is "A chef cooks more than most people", while an example of a false sentence is "A loud explosion can scare a tractor". Some sentences are a bit longer than others, but all of them intend to represent the same level of difficulty when it comes to comprehension. There are 37 items, but since children have just 2 minutes to go through the test, none of the participants in our sample managed to read all the sentences. The average length of the sentences is 6 words. The score is computed automatically by the app and equals the number of sentences judged correctly within the given time of 2 minutes.

Figure 1: The example of task in SV



Blå er en farge.

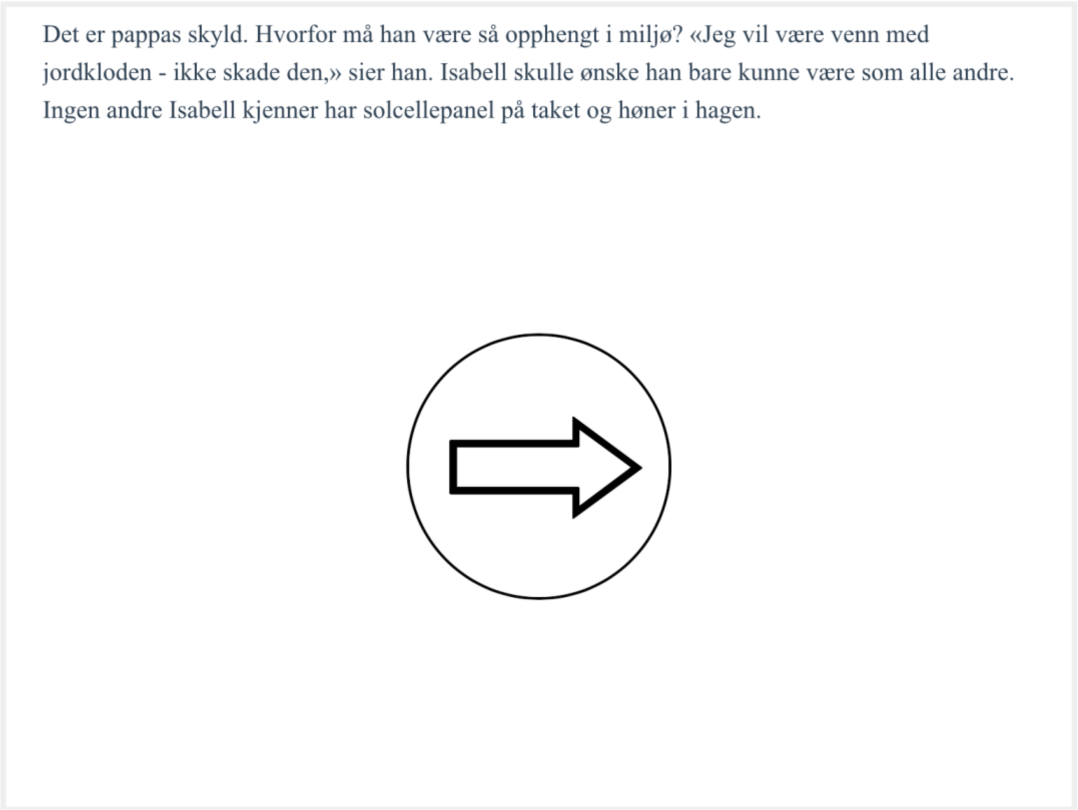
The test was administered individually as a part of a larger battery of tests used in the BetterReading project. Because children's decision and its execution (clicking on the chosen icon) are included in the measure, the participants were instructed to keep their hands close to the tablet to minimize differences in reaction time. The children were also informed to work with the task as fast as they can to maximize the chances of capturing their optimal reading rate by the measure. Before the test starts, the children had the possibility to practice solving the tasks with two trial exercises, which were not included in computing the final score.

Sentence Verification intends to measure silent reading comprehension efficiency at the sentence level. Since the emphasis is put on efficiency rather than on comprehension accuracy, the sentences were constructed in a way that ensure their accessibility. That is, they have a simple grammatical structure, and their content should be well known to children in fifth grade. This is likely to reduce demands on general knowledge and vocabulary. In the pilot study, sentences that were often read too slowly by participants or led to inaccurate responses were eliminated (Simonsen et al., 2022). Because SV measures fluency in relation to comprehension of unrelated, simple sentences, the test put very low demands on higher level text processing. The children do not have to make global inferences, integrate a lot of information, or construct a complex mental model of a longer passage. The test does not require either comprehension monitoring or a great capacity for working memory. Because of a short duration of the task, the demands on attention and stamina are also reduced. On the other hand, the scores on the test are probably influenced by the accuracy and the rate of decoding together with sentence level of comprehension (for example syntactic parsing and semantic integration). Simonsen and colleagues (2022) have defined reading comprehension efficiency as “the rate at which readers extract information from the text”, but the measure also includes the use of that information in the decision component and the motoric execution needed to carry out the task.

4.3.1.2 Picture Selection (PS)

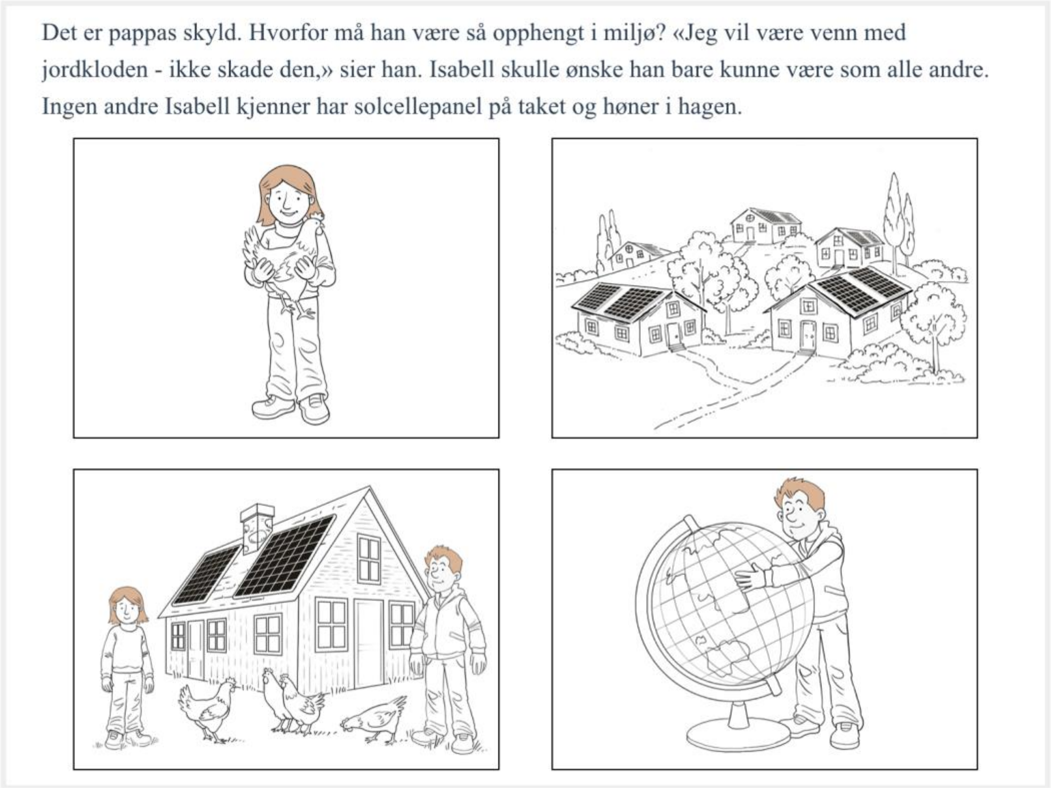
Picture Selection is an innovative test that was developed especially for the BetterReading project and implemented as an iPad app. In contrast to SV, which measures silence reading comprehension efficiency at the sentence level, Picture Selection measures comprehension at the passage level. The test uses short paragraphs that form a coherent story. The children´s task is to silently read paragraphs that appear on the screen one by one. After reading every passage, they are supposed to tap on an icon with an arrow that leads to the appearance of fourth, mostly black and white pictures (some details on a few pictures are colored). Next, the children must click on the picture that reflects in the best way the content of the passage. The passage is still visible over the four pictures, so children can reread the whole paragraph or parts of it if they want to monitor their comprehension. Visibility of the passage during the decision making might also reduce demands on memory. However, the task involves the visual processing of four different pictures that often contain different numbers of details, and will naturally pose some memory demands. Decoding skills, fluency, and higher-level text processing (inferencing within a paragraph, integrating information from the text with general knowledge, use of the knowledge about the text structure) probably impact the results of the test. Like in SV, the measure includes also time used to extracted information from the text and the time that is needed to use that information to solve the task. Scoring is computed automatically by the app and equals the number of correctly chosen pictures within the given time of 4 minutes.

Figure 2: An example of a task in PS -reading paragraph.



Det er pappas skyld. Hvorfor må han være så opphengt i miljø? «Jeg vil være venn med jordkloden - ikke skade den,» sier han. Isabell skulle ønske han bare kunne være som alle andre. Ingen andre Isabell kjenner har solcellepanel på taket og høner i hagen.

Figure 3: An example of a task in PS - a choice of a picture



Det er pappas skyld. Hvorfor må han være så opphengt i miljø? «Jeg vil være venn med jordkloden - ikke skade den,» sier han. Isabell skulle ønske han bare kunne være som alle andre. Ingen andre Isabell kjenner har solcellepanel på taket og høner i hagen.

The original story, which was written for BetterReading to be used in the PS test, tells about the everyday experiences of a girl, who is of the same age as the participants. The familiarity of the topic is supposed to secure the texts´ accessibility. Moreover, PS has also undergone pilot testing which resulted in eliminating items that were read too slowly or were too demanding in terms of comprehension. The average length of a passage is 41 words and there are 20 such items in the test. The time limit is 4 minutes, and none of the children in our sample has managed to read the story during this time. During test administration, the participants were instructed to work with the task as fast as they can, read all the passages before clicking on the arrow, and keep their hands close to the tablet. Before the test starts, the children practice solving the task with two practice trials. This is meant to ensure that children will understand the task before moving to items that are included in the final score.

4.3.1.3 Oral Reading Fluency (ORF)

ORF is a screener that intends to measure both components of oral reading fluency – accuracy in decoding and reading rate. Although the instrument also gives the possibility to evaluate prosody, that last construct was not assessed during the project. Under individual administration, the child is instructed to read texts as accurately and fast as possible.

The battery consists of three texts that have difficulty level adjusted to the age of participants. Misread or omitted words are considered mistakes. The person administering the test subtracts the number of mistakes from the number of all words the child has read in one minute. The instrument is based on the Dynamic Indicator of Basic Early Literary Skills (Good & Kaminski, 2002) and has shown good reliability and validity in assessing reading fluency in the Norwegian context (Arnesen et al., 2017).

4.3.1.4 Neale Analysis of Reading Ability (NARA)

NARA is an assessment instrument that intends to measure reading comprehension on a text level, along with oral reading rate, and accuracy. The original test was developed by Neale (1958) and the BetterReading has used its´ Norwegian version. Testers in the BetterReading project have noted the time children have used to read each text and the number of misread words; however, this thesis will use only information about participants´ comprehension accuracy. The children´s task was to read the texts from the battery and answer open-ended comprehension questions after each text. Before reading, children were also presented briefly with a picture that gave them a hint about the content of the text. Prior to administration of items that were included in computing the score, the children practiced the task on one short text. Some of the questions required memory for details, for example, the name of a character. To answer some other questions children needed to integrate different pieces of information from the text and their general knowledge. The test puts also substantial demands on the use of expressive language (Cain & Oakhill, 2006). There were 6 different texts with increasing difficulty levels in terms of letter size, text length, grammatical structures, and vocabulary. Assessment is discontinued if children do not manage to answer any question from a text. The test has English norms, but they were not used in the study.

4.3.1.5 Test of Word Reading Efficiency (TOWRE)

TOWRE was developed by Torgesen and colleagues (2012), and it intends to assess children´s word-level reading skills. The BetterReading research group has used Norwegian digital version of the test in the project. The test has two subtests, and each of them contains two parts – A and B. The first subtest "words" intends to assess children´s sight word vocabulary, and it indicates how good the participants are at recognizing single words effortlessly and quickly. The purpose of the second subtest "pseudowords" is to measure children´s skill to use their graphophonetic knowledge to read unknown words. The participant´s task is to read words or pseudowords from each part of the test as fast and as accurately as it is possible. The children have only 45 seconds to complete each task. The items are presented in four vertical columns in each part of the test. While children read the items on paper, the tester scores the items on a tablet. Testers were provided with a scoring guide that gave examples and rules for the acceptable pronunciation of pseudowords. Additionally, testers could listen to the recording from the assessment to check the scores later. Before administration of each subset the children read word or pseudoword lists as a short practice exercise.

The manual of TOWRE-2, which was developed by Torgesen and colleagues, provides data that indicates good validity and reliability of the instrument. It is a popular measure of word and pseudoword oral reading fluency in research, schools, and diagnostic practice in many parts of the world (Tartar et al., 2015).

4.3.1.6 Discrete Words

Discrete Words is a digital test developed for the BetterReading project. The instrument intends to assess how fast and accurately children read single words. The words appear in the middle of the screen of a laptop one by one. After children read a word, the tester clicks on the mouse, and the new word appears. Children are instructed to read words just once and to not correct themselves. The program contains some words that are used in intervention materials and some others that have the same difficulty level. This thesis will use just scores computed with items that are not included in intervention program. The items have a similar frequency in the Norwegian language, orthographic complexity, morphologic and syllabic structure. They are displayed in random order. The program measures the time children spend reading each word, and the accuracy of word reading is manually scored later. Therefore, the results give two variables: the rate and the accuracy of oral reading of single words. Prior to administration of items that are included in the final scores, children are presented with three practice trials.

4.3.1 Measures used only with the Intervention Group

4.3.2.1 The Norwegian Version of The British Vocabulary Scale (BPVS II)

BPVS intends to measure receptive vocabulary, which is defined as the understanding of words. The Norwegian version of the test is standardized and has age norms for children from 3 to 16 years. 884 children from all parts of Norway took part in the standardization process. The results of the validation study indicated good reliability and validity of the instrument. The correlation analyses confirmed also theoretical assumptions about the association of grammatical skills and vocabulary in various age groups.

The BetterReading group has used computerized version of the test. BPVS comprises 12 sets, and each of them includes 12 items. During individual assessment, a word was presented auditorily by the program and the children´s task was to choose the picture that represents the word’s meaning. The child signalized the choice by tapping the picture. Therefore, the test does not put high demands on active language use. The item´s difficulty increasesed from one set to the next. The children started with the words that are assigned for their age. When the participant made more than one mistake in a set, the words from a previous set for younger children were administered. The child had to go through all the items in each set, but the test stopped when the child made eight or more mistakes. For each correct answer, the child gots one point, and the final score is the sum of the points (Lyster et al., 2010).

4.3.2.2 Test of Reception of Grammar (TROG)

The Test of Reception of Grammar was developed by Dorothy Bishop and adapted in the Norwegian context by Lyster and Horn (2009). During the norming process, 950 participants were tested and the instrument has currently norms for ages from 4 to 16 years. The test intends to assess receptive grammatical understanding. 20 blocks help to examine children´s perception of inflection, word order, and word function. During the individual assessment, a sentence is presented auditorily for each item, for example: “The sheep runs”. The children´s task is to point at one of four pictures that represents the content of the sentence. The pictures are very simple, but it can not be precluded that children´s visual attention can impact the results. The task, however, does not put demands on expressive language – the active use of the language to convey the meaning. Each block consists of four items. The score is the sum of the points from all the blocks (Lyster & Horn, 2009).

4.3.2.3 RAVEN

RAVEN is a test that intends to measure non-verbal intelligence. The children´s task is to look at color matrices that are displayed on a sheet of a paper or on screen, and then find out the system that determines how the geometric figures are arranged in the matrix. Next, the children must look at different versions of “bricks” below the matrix. The bricks contain different geometric figures, and the children are supposed to point to the one that fits the matrix above. The instrument taps different skills, like visual attention, logical thinking, working memory, and spatial and categorization ability. The test does not have a time limit. The test can be administered individually or in a group (Helland-Riise & Martinussen, 2017).

4.4 Reliability

Reliability is a prerequisite for validity, and it indicates the extent to which an assessment instrument gives consistent and reproducible scores reflecting a relatively stable skill or feature of a tested person. According to classical Test Theory, each measurement results in the combination of a true score, which contains information about participants’ real standing on a particular attribute, and measurement error. The latter is a result of random factors, such as test situation, or participant features (which are often temporary, such as exhaustion) not connected to the construct that the test intends to assess. Measurement error can have a positive or negative effect on the participants´ score (John & Benet-Martinez, 2014).

In this study, the test-retest method was used to assess the stability of Picture Selection and Sentence Verification. The tests were administrated two times in the intervention group, and the correlation between scores from these two timepoints will be computed. There was a two-month interval between the two test points. Therefore, we hypothesize that the possible improvement of the scores is due to the *reactivity effect* or *carry-over effect*. The former involves improvement of test results due to familiarization with the test situation and the demands of the test, while the latter is a consequence of participants’ memory for some of their previous answers. Different levels of required stability are found in the literature. While standardized tests should have a reliability coefficient of .80 or higher, for the purpose of screening a reliability coefficient of .70 or higher is often considered acceptable (Murphy & Davidshofer, 1994).

Another form of reliability is internal consistency, that concerns the extent to which different items in a test assess the same construct (John & Benet-Martinez, 2014). In the thesis, internal consistency of NARA, Discrete Words (accuracy), RAVEN, TROG, BPVS, and ORF are reported.

4.5 Construct Validity

While classical Test Theory states that measures do not perfectly reflect the skill of an assessed person because of measurement error, more modern approaches acknowledge that the difficulty of accurate operationalization of a construct can also lead to systematic error embedded in the instrument (John & Benet-Martinez, 2014). The operationalization of an abstract construct involves choosing its measurable indicators (Lund, 2002). In psychology, as well as in special needs education, there is rarely one universal way to operationalize not directly observable phenomena, and there is a threat that scores may be impacted by other skills that are not included in the theoretical construct.

The examination of construct validity aims to assess to which extent the instrument is measuring the skill it is supposed to. It involves both reliability checks and investigation of convergent and divergent validity. Evidence of convergent validity can be gathered by checking associations with other, already validated tests, that intend to measure the same or similar construct. In contrast, the investigation of divergent validity requires the examination of associations with measures that should not be strongly related to the theoretical construct (John & Benet-Martinez, 2014).

In this study, silent reading comprehension efficiency is conceptualized as a reading skill and that is why scores on PS and SV should correlate with other tests that measure reading skills, but not with non-verbal intelligence. Although there is evidence that receptive vocabulary and grammar are associated with reading comprehension (Cain, 2010), the results on PS and SV will probably not be strongly correlated with them, because SRCE also requires higher level language skills, during a reading of an accessible text, than grammar and vocabulary.

Because PS assesses comprehension on a passage level, it is assumed that this test will correlate more highly with the comprehension accuracy test – Neale Analysis of Reading Ability (NARA) – than SV. The notion of SRCE involves the effective silent reading of connected text, and thus it is possible that both tests developed by BetterReading will correlate more strongly with Oral Reading Fluency (ORF), which examines fluency during an oral reading of a passage, than with tests that measure reading of single, unrelated words.

4.6 Statistical Validity

Statistical validity concerns whether it is justified to make inferences about the relation between two variables. An association between variables that is not reasonably strong and statistically significant does not allow to draw conclusions about investigated phenomena. In this case, observed relations could be a result of coincidence or bias that occurred during sampling. Therefore, statistical validity is perceived as a prerequisite for any other types of validity.

Statistical power depends on sample size, variability of the skill in the population, the magnitude of the effect of the variable, and choice of significance level (Lund, 2002). In this thesis, the desired significance level is .05, which implies that correlations with p-value less than .05 will be considered as significant.

Statistical testing of hypothesis always involves a risk of errors, which are traditionally divided into Type-I error and Type-II error. The first one occurs when the null hypothesis about the lack of a relationship between variables is rejected, while the null hypothesis is true. The latter error is made when the null hypothesis is accepted but in fact, there are relationships between variables. With a significance level of .05, there is 5% probability of making the Type-I error (Navarro & Foxcroft, 2019)

4.7 External Validity

External validity concerns the extent to which we can generalize findings from a study to the "real world" – to a broader population of interest. Moreover, examination of external validity also requires evaluation if conclusions drawn from the study are applicable across different times and situations. One of the threats to external validity is selection bias, which occurs when the sample in the study is not representative for a wider population. The lack of representativeness could be a result of the homogeneity of individuals within a sample, small sample size, or not randomized sampling (Lund, 2002).

The main threat to external validity in this study is the use of a convenience sample, which means that only children that were easy for researchers to access had a chance to participate in the study. Moreover, because of ethical considerations only pupils that volunteered to take part in the project, were included in the sample. Furthermore, the thesis attempts to validate PS and SV and examine their potential to be used as screeners in Norwegian schools, but only children from Oslo-area who attend Norwegian school for 5 years were assessed with these instruments.

4.8 Ethical Considerations

The research project BetterReading has received approval from *the Norwegian Center for Research Data* (NSD). This thesis is a part of the project, and it follows the requirements described in *The Guidelines for Research Ethics* (NESH, 2016).

In addition to the collection of written consent from the parents of children who volunteered to take part in the study, the pupils were asked if they agree to participate in the project. The parents as well as children have received information about the study and the possibility to withdraw the consent and leave the project at any time. The information about the study and the participants' rights were adapted to the children´s age. The data was anonymized by replacing pupils´ names and the names of the school with ID-number. The participants received their ID-numbers at the beginning of the assessment situation and there was no list that could make it possible to connect the names with ID-codes. The master students and other research assistants have received training, which guided them on how make the assessment a pleasure and stress-free situation for children. additionally, all the test leaders were instructed on how to treat data with respect to the principles of privacy, anonymity, and confidentiality. Each child has received a diploma and stickers as appreciation for their participation.

5. Results

5.1 Descriptive Statistics of Variables

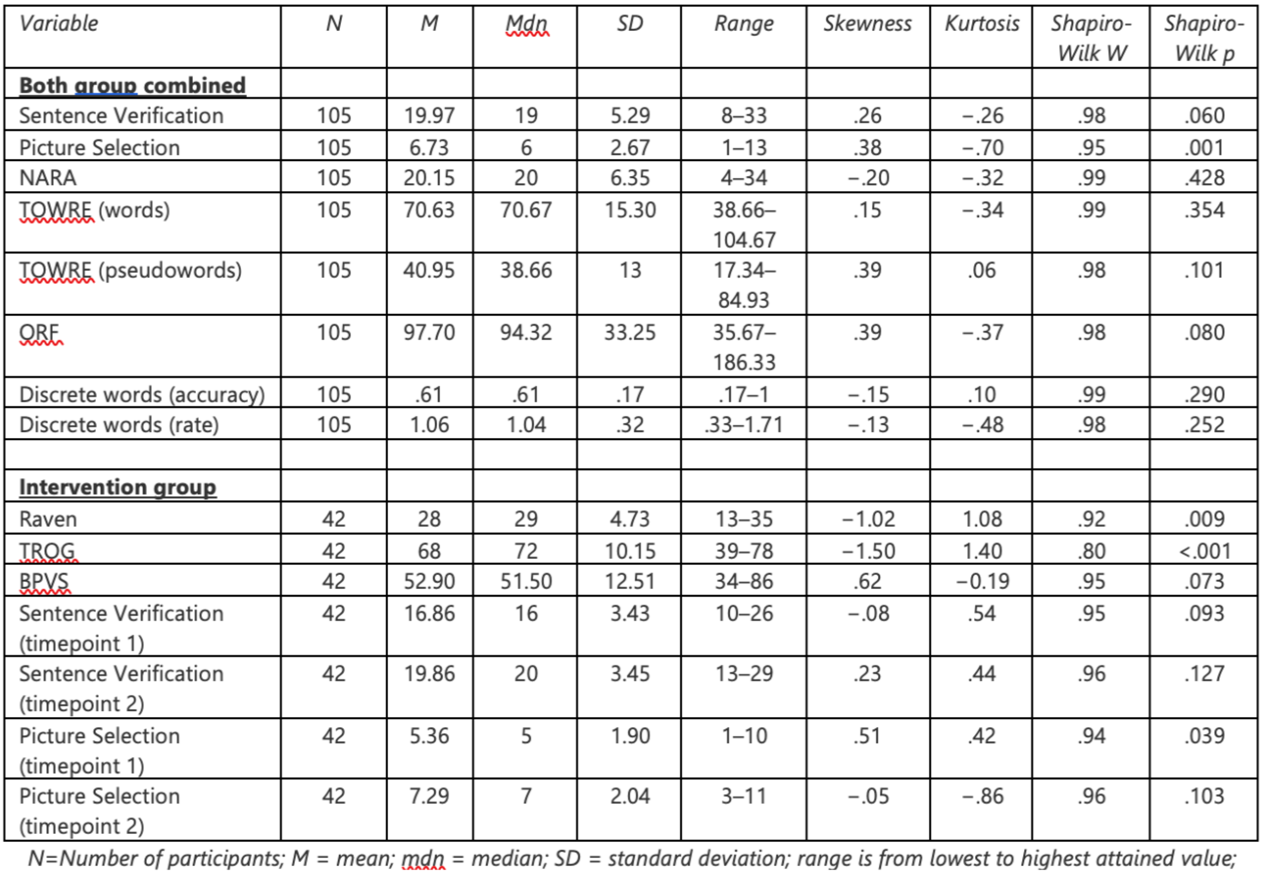
The study´s sample consisted originally of 111 pupils. However, 6 cases were filtered out from the sample due to missing data or procedural mistakes during the assessment. For example, some of the children misunderstood instructions or could not finish the tasks. The current sample comprises 105 cases (46 intervention children and 59 non-intervention children). Only 46 children were tested with RAVEN, BPVS, TROG as well as PS and SV. Due to missing data only 42 of them will be included in the analysis of test-retest reliability. All 105 children was tested with ORF, but the intervention group has read just one of the three texts from that battery. Therefore, their scores are not included in internal consistency analysis. Additionally, one child in the non-intervention group did not read 2 of the texts and was removed from the analysis of internal consistency.

Firstly, the descriptive statistics and histograms of each variable will be presented to summarize the collected data. The mean shows the average value of the data and is most appropriate to use with numerical data that is normally distributed because it is quite sensitive to extreme values. When the data do not meet assumptions of normal distribution, it may be more informative to use the median, which shows the middle value in the sample. The normality of distribution will be examined by conducting the Shapiro-Wilk test, evaluating the histogram, and calculating skewness and kurtosis. Skewness gives information about asymmetry of distribution. When skewness equals 0, it indicates perfect symmetry, whereas negative values suggest that data have a lot of small values, and the distribution is “left skewed”. On the other hand, data that is right skewed have a lot of values that are greater than the mean and skewness takes positive values (Navarro & Foxcroft, 2019).

Another descriptive statistic that helps in evaluating normality is kurtosis. It gives information about how many data points are far from the mean. Negative kurtosis indicates that there are too few data points from the mean, and positive kurtosis suggests that there are too many data points far from the mean. However, data with skewness and kurtosis that do not exceed 1 (or −1) is usually considered as approximately normally distributed (Cohen & Swerdlik, 2018).

For tests that are not time limited internal consistency will be also presented, which is a type of reliability. It shows to which extent different items of an assessment instrument are consistent with each other. In other words, internal consistency indicates if different questions or tasks from a test assess the same construct (Punch & Oancea, 2014). Internal consistency will be evaluated with Cronbach´s alpha (α) and McDonald´s omega (ω). Values of omega and alpha that are equal to or higher than .70 are considered as indicators of good internal consistency (Navarro & Foxcroft, 2019).

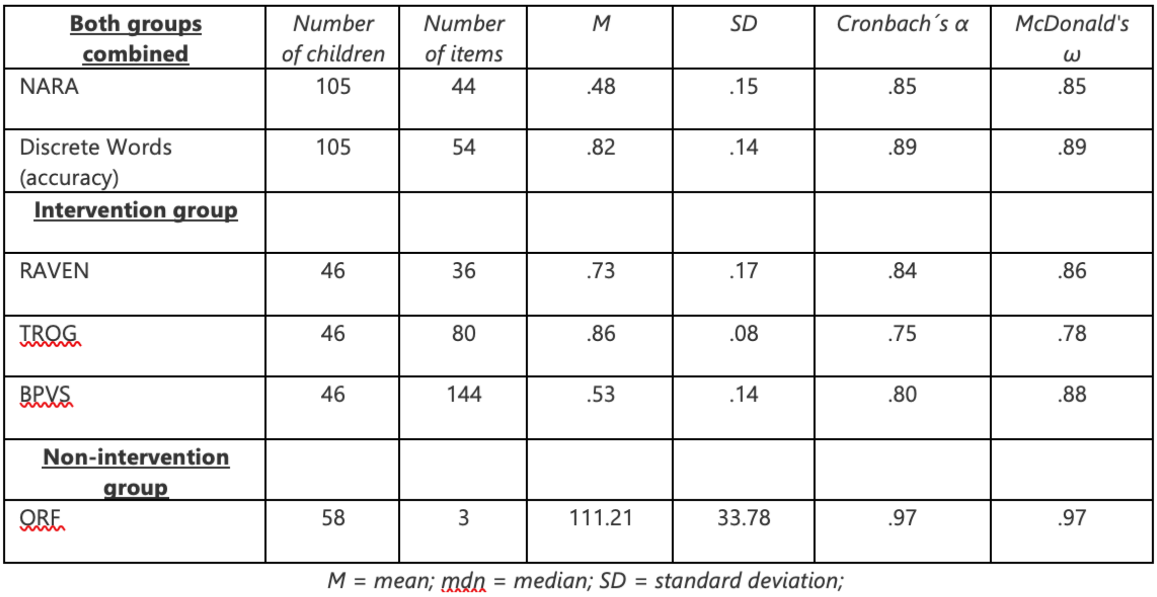
Table 1: Descriptive statistics of variables – 1



Variable	*N*	*M*	*Mdn*	*SD*	*Range*	*Skewness*	*Kurtosis*	*Shapiro-Wilk W*	*Shapiro-Wilk p*
Both group combined									
Sentence Verification	105	19.97	19	5.29	8–33	.26	−.26	.98	.060
Picture Selection	105	6.73	6	2.67	1–13	.38	−.70	.95	.001
NARA	105	20.15	20	6.35	4–34	−.20	−.32	.99	.428
TOWRE (words)	105	70.63	70.67	15.30	38.66–104.67	.15	−.34	.99	.354
TOWRE (pseudowords)	105	40.95	38.66	13	17.34–84.93	.39	.06	.98	.101
ORF	105	97.70	94.32	33.25	35.67–186.33	.39	−.37	.98	.080
Discrete words (accuracy)	105	.61	.61	.17	.17–1	−.15	.10	.99	.290
Discrete words (rate)	105	1.06	1.04	.32	.33–1.71	−.13	−.48	.98	.252
Intervention group									
Raven	42	28	29	4.73	13–35	−1.02	1.08	.92	.009
TROG	42	68	72	10.15	39–78	−1.50	1.40	.80	<.001
BPVS	42	52.90	51.50	12.51	34–86	.62	−0.19	.95	.073
Sentence Verification (timepoint 1)	42	16.86	16	3.43	10–26	−.08	.54	.95	.093
Sentence Verification (timepoint 2)	42	19.86	20	3.45	13–29	.23	.44	.96	.127
Picture Selection (timepoint 1)	42	5.36	5	1.90	1–10	.51	.42	.94	.039
Picture Selection (timepoint 2)	42	7.29	7	2.04	3–11	−.05	−.86	.96	.103

N=Number of participants; M = mean; mdn = median; SD = standard deviation; range is from lowest to highest attained value;

Table 2: Descriptive statistics of variables - 2: Internal consistency



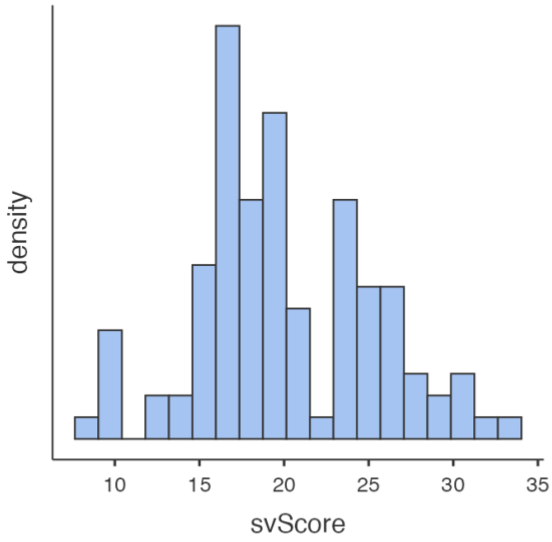
Both groups combined	*Number of children*	*Number of items*	*M*	*SD*	*Cronbach´s α*	*McDonald's ω*
NARA	105	44	.48	.15	.85	.85
Discrete Words (accuracy)	105	54	.82	.14	.89	.89
Intervention group						
RAVEN	46	36	.73	.17	.84	.86
TROG	46	80	.86	.08	.75	.78
BPVS	46	144	.53	.14	.80	.88
Non-intervention group						
ORF	58	3	111.21	33.78	.97	.97

M = mean; mdn = median; SD = standard deviation;

5.1.1. Sentence Verification (both groups combined)

The histogram shows quite symmetric distribution with small right skewness of .26 and kurtosis of -.26. There is one prominent peak about value of 17, but there are two noticeable gaps – one on the left side, and another closer to the center of distribution. The mean is 19.97 and the median is 19. The variable is distributed over a range from 8 to 33. The results from Shapiro-Wilk test do not allow to reject the null hypothesis that the scores are sampled from a normal distribution, therefore the variable will be considered further as normally distributed.

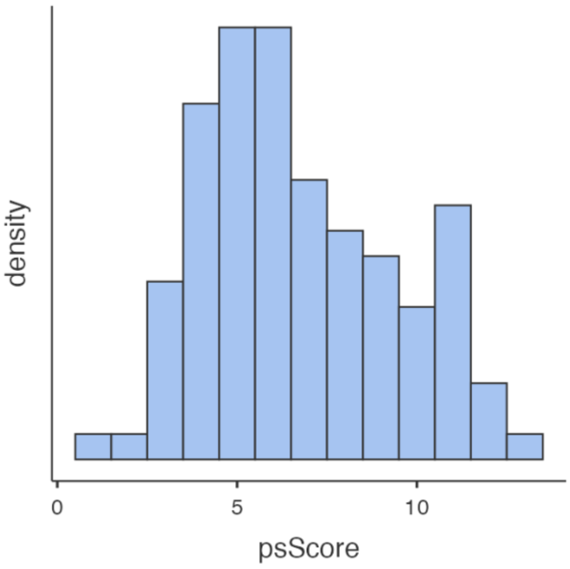
Figure 4: Histogram of Sentence Verification (both groups combined)



5.1.2 Picture Selection (both groups combined)

The results of the Shapiro-Wilk test indicate that scores from Picture Selection attained by the non-intervention group should not be considered as normally distributed. The median equals 6, and the mean is 6.73. Reported standard deviation has to be also used with caution. The distribution has no clear peak and is right skewed with the skewness of 0.38. Kurtosis of -.70 reflects few values on the tails of the distribution. The variable is distributed over a range from 1 to 13.

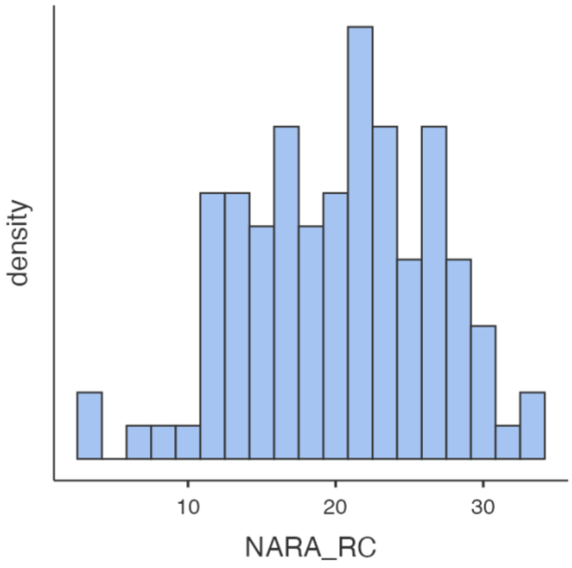
Figure 5: Histogram of Picture Selection (both groups combined)



5.1.3 NARA (both groups)

The histogram shows that NARA has a distribution with a slight left skew of -,20, while kurtosis equals -,32. Both values suggest that the data does not extensively deviate from normality. The results from the Shapiro-Wilk test also do not allow to reject the null hypothesis that the scores are sampled from a normal distribution, therefore the variable will be considered further as normal distributed. The mean is 20.15 and the median is 20. The range of scores lies between 4 and 34 points. The test has an internal consistency of α = .85, and ω = .85.

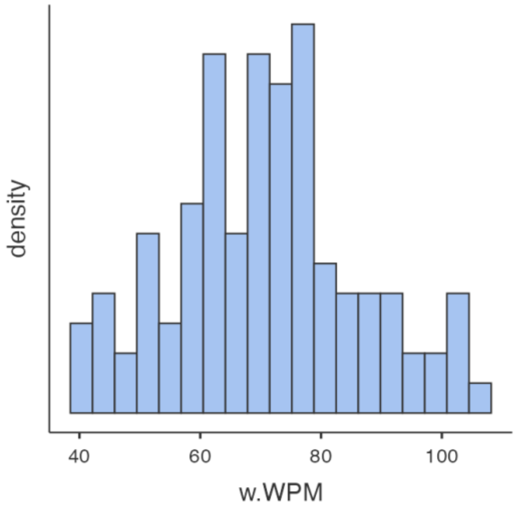
Figure 6: Histogram of NARA - both groups combined



5.1.4 TOWRE – words (both groups combined)

Scores from the first part of TOWRE which assesses the decoding of words form a slightly right skewed distribution with a prominent peak around 79. The skewness of .15 and kurtosis of -.34 indicates that the distribution is approximate to normality. The Shapiro-Wilk test does not permit to reject of the null hypothesis that the scores are sampled from a normal distribution, therefore the variable will be considered further as normally distributed. The scores take values from 38.66 to 104.67 which indicates big individual differences in assessed skill.

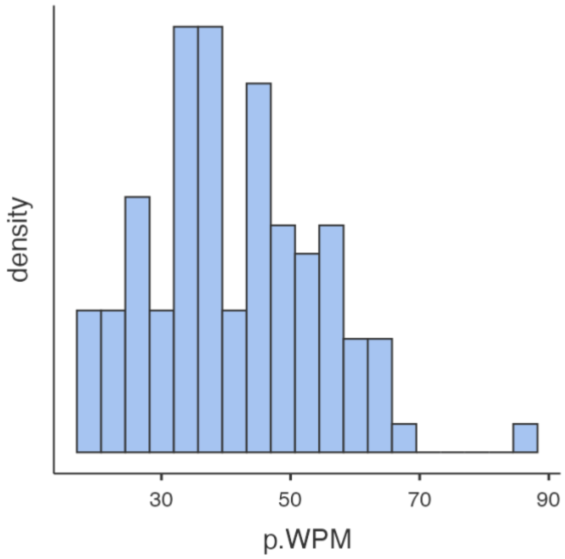
Figure 7: Histogram of TOWRE-words (both groups combined)



5.1.5 TOWRE – pseudowords (both groups combined)

The scores from the second part of TOWRE, which intend to assess decoding of pseudowords, form a slightly right skewed distribution with skewness of .39. The kurtosis is quite small and equals .06. The score of 84.93 is the highest one and seems to be outlier because it lies 18.93 points below next highest score, which is 66.00. The minimal score is 17.34, while the mean and median equals respectively 40.95 and 38.66. There is no one clear peak. Histogram suggests a adequate approximation to normal distribution. Moreover, the results from the Shapiro-Wilk do not allow to reject the null hypothesis that the scores are sampled from a normal distribution, therefore the variable will be considered further as normal distributed.

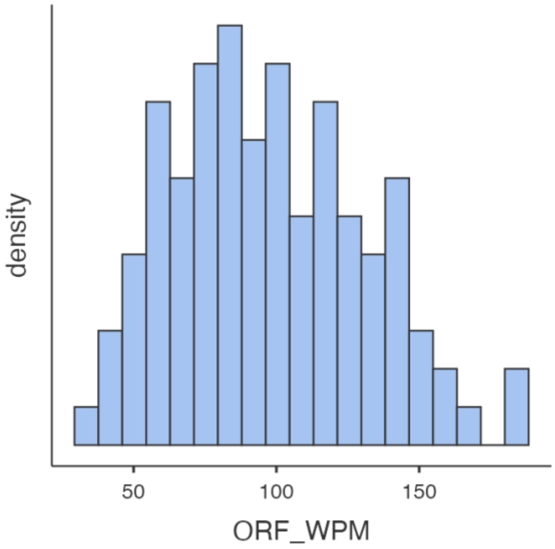
Figure 8: Histogram of TOWRE - pseudowords (both groups combined)



5.1.6 ORF (non-intervention)

The mean of the scores from ORF is 97.70, while the median is 94.32. There is a big spreading in scores from 35.67 to 186.33 points, which suggests that there were big individual differences in children´s performance on that test. The distribution is right skewed with a skewness of .39 which indicates more lower values than higher ones. Kurtosis of -.37 suggests that there are fewer data points on the ends of the distribution than would be expected from a perfectly normal distribution. However, the histogram indicates good approximation to normal distribution. The results do not allow for rejecting the null hypothesis that the scores are sampled from a normal distribution; therefore, the variable will be considered further as normal distributed. The test has an internal consistency of α = .97, and ω = .97.

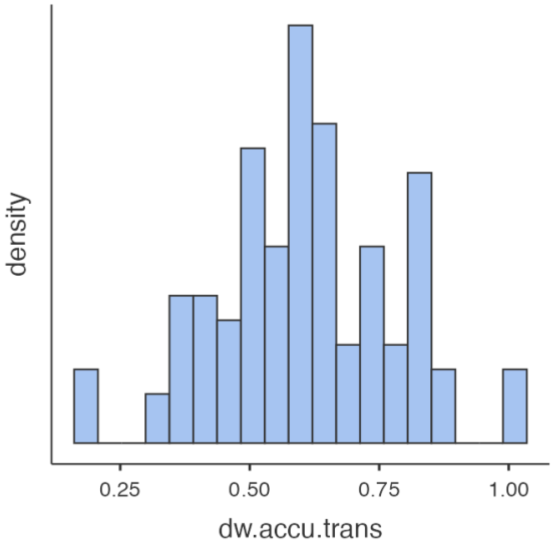
Figure 9: Histogram of ORF - non-intervention group



5.1.7. Discrete Words – Accuracy (both groups combined)

The range of transformed scores from Discrete Words (accuracy) spreads from .17 to 1.00, while both mean, and median is .61. The distribution with left skewness of -.15 and kurtosis of .10 approximates well the normal distribution. The Shapiro-Wilk test does not allow to reject the null hypothesis that the scores are sampled from a normal distribution, therefore the variable will be considered further as normal distributed. Histogram shows two gaps – one on the right end of the distribution, and one on the left. The test has an internal consistency of α = .89, and ω = .89.

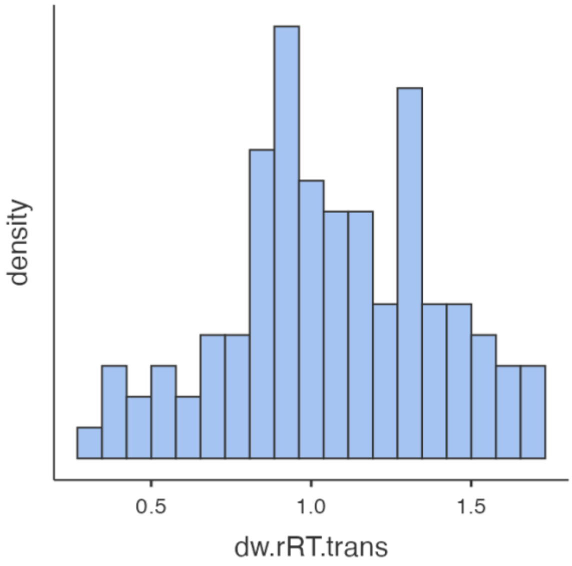
Figure 10: Histogram of Discrete Words - accuracy (both groups combined)



5.1.8 Discrete Words – Rate (both groups combined)

The histogram shows two clear peaks – one close to the center, and one on the right side of distribution. However, the small value of left skewness, -.13, approximates collected data to symmetric distribution. Kurtosis of -.48 is reflected on the histogram by a few data points on the ends of the tails. The absolute value of kurtosis does not exceed 1 which indicates that data does not deviate too strongly from normal distribution. Additionally, the Shapiro-Wilk test does not allow to rejection the null hypothesis that the scores are sampled from a normal distribution, therefore the variable will be considered further as normal distributed. The mean of 1.06 is close to the median which equals 1.04. The transformed scores are spread between .33 and 1.71 words per second.

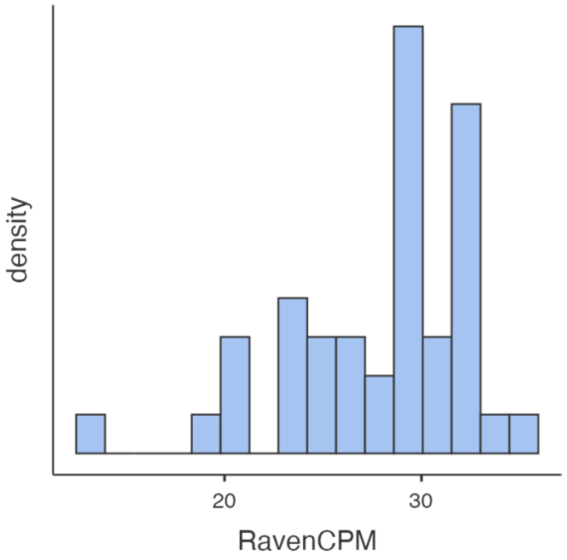
Figure 11: Histogram of Discrete Words - rate (both groups combined)



5.1.9. RAVEN (intervention group)

The histogram shows the asymmetrical distribution of scores from RAVEN. The left skewness of -1.02 indicates a strong deviation from the normal distribution and suggests that the majority of children has attained high scores on the test. The kurtosis of 1.08 reflects some data points that are visible on the end of the right tail, while there are 2 gaps on the left tail. There are two prominent peaks of about 29 and 33 on the right tail. The Shapiro-Wilk test allows for rejecting the null hypothesis that the data is sampled from a normal distribution, therefore the variable will be considered further not normally distributed. The median equals 29, while the maen is 28. Reported standard deviation must be also used with caution. The highest score is 35, and the lowest score is 13 and it is probably an outlier. The test has an internal consistency of α = .84, and ω = .86.

Figure 12: Histogram of RAVEN - intervention group

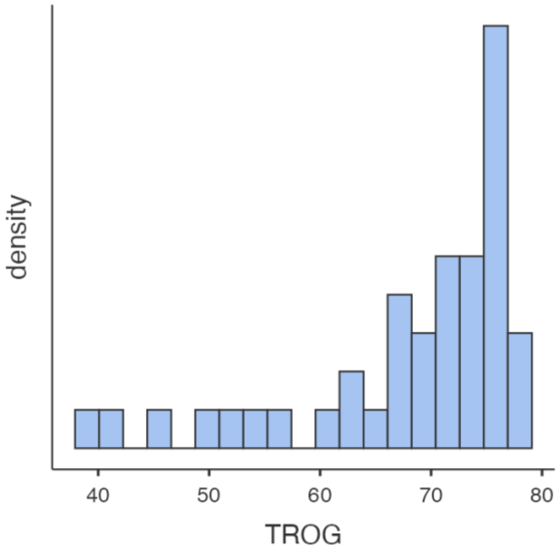


5.1.10 TROG (intervention group)

Although the range of scores from TROG lies between 39 and 78, the distribution is extensively left skewed, and most of the values are clustered towards the end of the right tail. Both the mean of 68 and the median of 72 are pulled towards the right tail.

The skewness is -1.50 and the kurtosis is 1.40. It indicates that data deviates strongly from normality and that most of the children have performed well on the task. That is aligned with results from the Shapiro-Wilk test which allows for rejecting of the null hypotheses that the data is sampled from a normal distribution. Therefore, the variable will be considered further as not normally distributed. The test has an internal consistency of α = .75, and ω = .78.

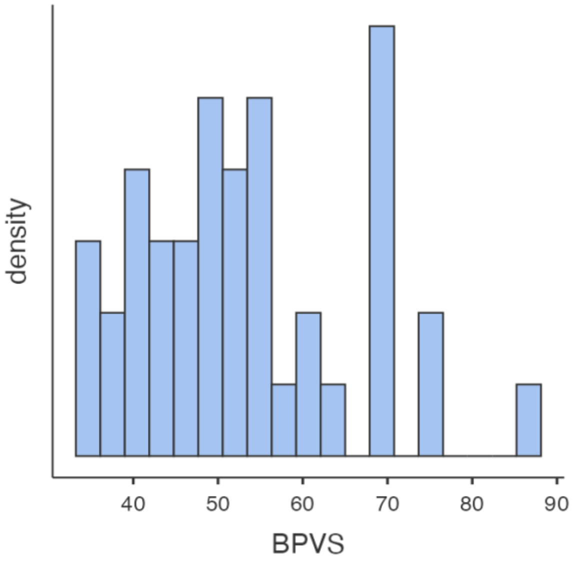
Figure 13: Histogram of TROG - intervention group



5.1.11 BPVS (intervention group)

The distribution of scores from BPVS has right skewness of .62 which reflects that most of the data points are clustered on the left side of distribution. However, the humped peak of about 70 is placed on the right side from both the mean (52.90) and the median (51.50). The minimum score is 34. Kurtosis of -0.19 reflects a little bit more data points on the left side than it is usually expected from a perfectly normal distribution. However, the Shapiro-Wilk test does not allow for rejecting of the null hypotheses that the data is sampled from a normal distribution. Therefore, the variable will be considered further as normally distributed. The test has an internal consistency of α = .80, and ω = .88.

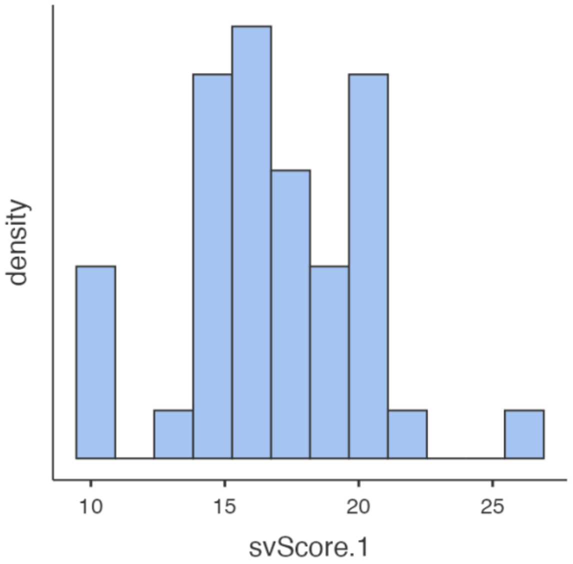
Figure 14: Histogram of BPVS - intervention group



5.1.12 Sentence Verification – timepoint 1 (intervention group)

The histogram shows a relatively symmetric distribution with small left skewness of -.08. In the same time kurtosis of .54 indicates that there are more data points on the ends of both tails that it could be expected from a perfectly normal distribution. The distribution has two gaps and one distinct peak of about 16. The scores are distributed over the range from 10 to 26. The Shapiro-Wilk test do not allow for rejecting of the null hypotheses that the data is sampled from a normal distribution. The data will be considered further as normally distributed. The median is quite close to the mean of 16.86 and equals 16.

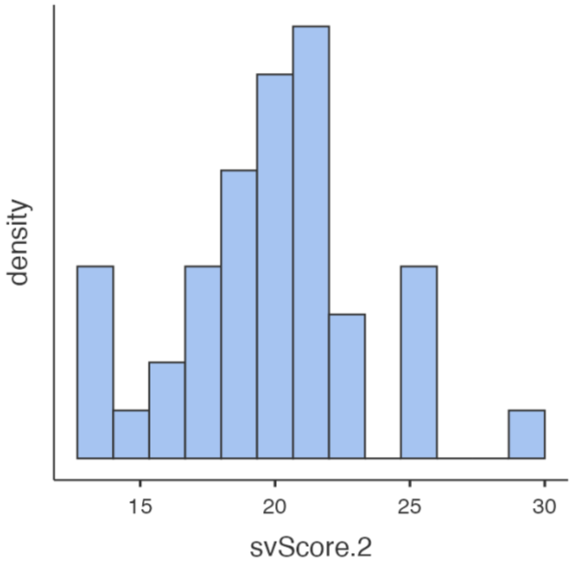
Figure 15: Histogram of Sentence Verification – timepoint 1 (intervention group)



5.1.13 SENTENCE VERIFICATION – timepoint 2 (intervention group)

The range of the scores from timepoint 2 is spread between 13 and 29 points, while the mean is 19.86 and the median equals 20. All these values indicate a slight improvement in children´s performance in relation to assessment from timepoint 1. The distribution has two gaps, and it is less symmetrical with right skewness of .23. Both histogram and kurtosis of .44. indicates that there are more data points on the ends of tails (especially on the left tail) than there is expected from a perfectly normal distribution. However, the Shapiro-Wilk test do not allow for rejecting the null hypothesis that the data is sampled from a normal distribution. Therefore, data will be considered as normally distributed.

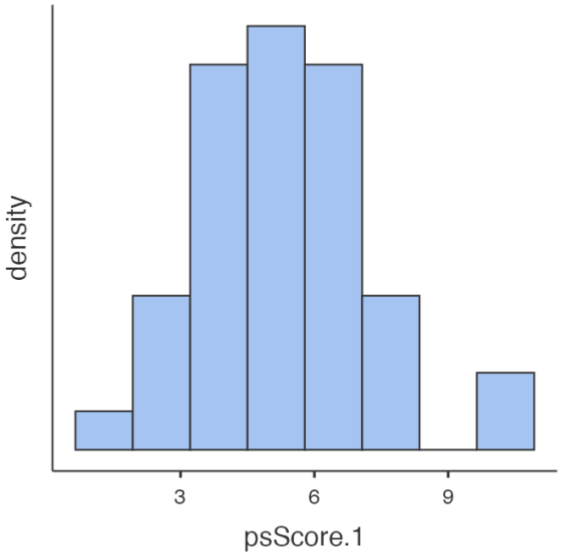
Figure 16: Histogram of Sentence Verification – timepoint 2 (intervention group)



5.1.14 PICTURE SELECTION - timepoint 1 (intervention group)

The scores with range between 1 and 10 form distribution with one peak of about 5 and with one gap on the right tail. The distribution has a right skew of .51 and kurtosis of .42. The Shapiro-Wilk test allows for rejecting of the null hypotheses that the data is sampled from a normal distribution. The data will be considered further as not normally distributed. The median is 5 and the mean equals 5.36.

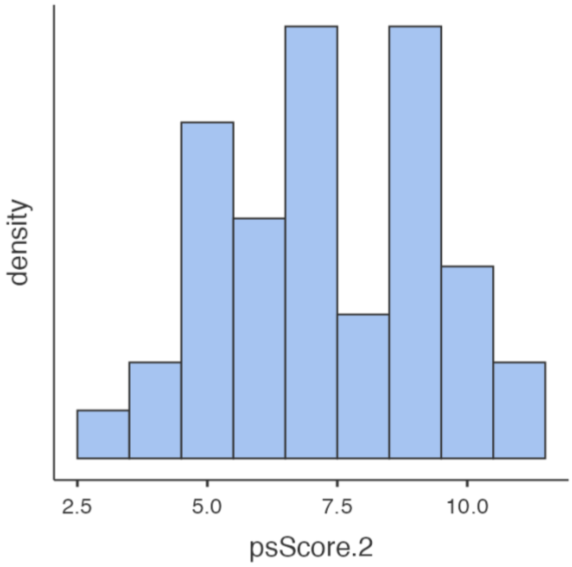
Figure 17: Histogram of Picture Selection - timepoint 1 (intervention group)



5.1.15 PICTURE SELECTION – timepoint 2 (intervention group)

The distribution of scores has a little left skew of -.05. However, kurtosis of -.86 indicates that there are more data points on the ends of the distribution that is expected from a perfectly normal distribution. Despite that, the Shapiro-Wilk test do not allow for rejecting of the null hypotheses that the data is sampled from a normal distribution. Therefore, data will be considered as normally distributed. The scores are spread between 3 and 11 points, the mean is 7.29 and the median equals 7. The higher values of the minimal and maximal score, together with the change of the mean and the median in relation to timepoint 1 indicates slight improvement of performance on that task.

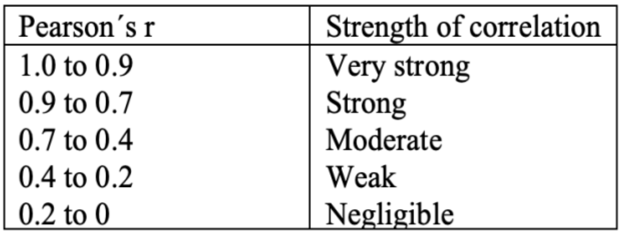
Figure 18: Histogram of Picture Selection – timepoint 2 (intervention group)



5.2 Bivariate Correlations

To answer research questions about validity, divergent validity and test-retest reliability of PS and SV correlation analysis will be conducted. The analysis intends to examine if two variables vary systematically, and what is the strength of the relation between them. Pearson´s *r* can take values from −1 to +1 and indicates what is the strength and direction of the association. The direction, however, does not involve causality. When r is negative, it simply means that the values of one variable increase, while the values of the other decrease. In contrast, the positive value of *r* suggests that when the values of one variable increase, the same happens to the second one. *r* that equals 0, however, suggests that there is no relationship between variables (Cohen et al., 2018). There are different ways to evaluate the strength of correlation in literature, but in this thesis, I will use the guide proposed by Navarro and Foxcroft that are displayed in table number 3 (Navarro & Foxcroft, 2019).

Table 3: Pearson's r and characterization of correlation´s strength



Pearson´s r	Strength of correlation
1.0 to 0.9	Very strong
0.9 to 0.7	Strong
0.7 to 0.4	Moderate
0.4 to 0.2	Weak
0.2 to 0	Negligible

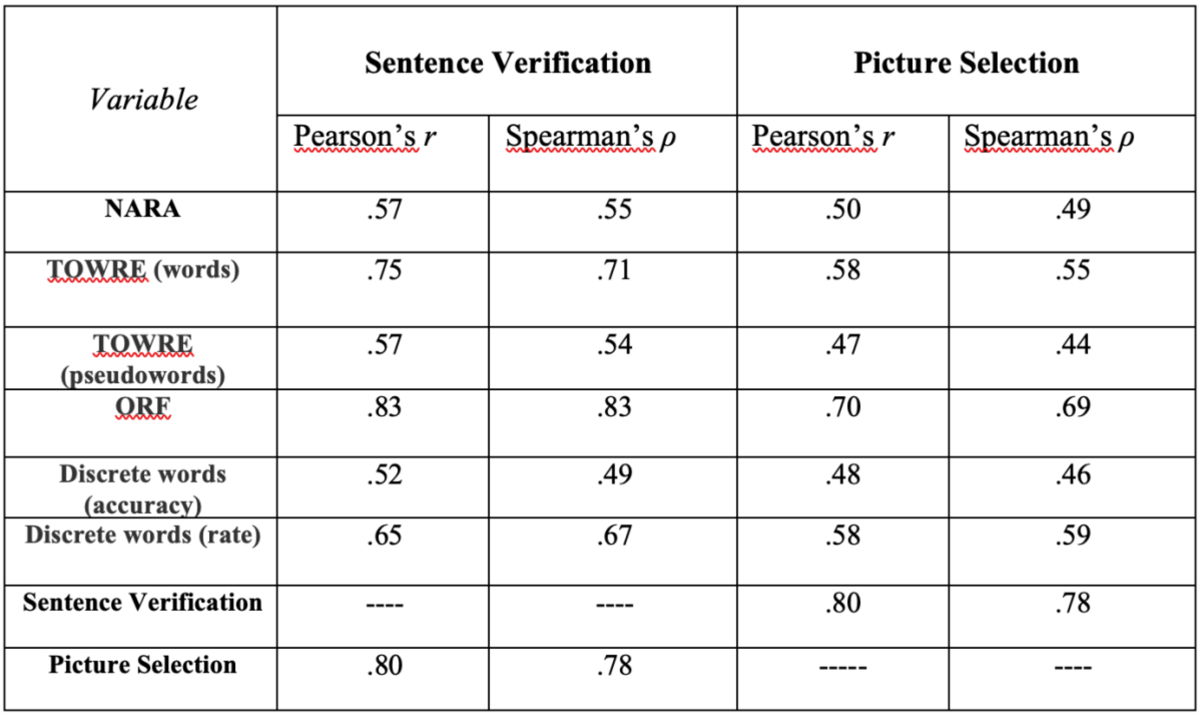
Person´s r is used for estimating correlation between normally distributed variables that are linearly related. However, the results from the Shapiro-Wilk tests and the analysis of graphs from descriptive statistics showed that not all variables of interest can be considered as normally distributed. Therefore, the correlations will be also reported with Spearman's *ρ,* which is appropriate to use in cases of non-normal distributions and monotonic correlations. Values of Spearman's *ρ* range also from -1 to 1, and this rank correlation coefficient is similar in interpretation to Pearson´s r – the closer Spearman's *ρ is to* – 1 or + 1, the stronger is the correlation between variables (Navarro & Foxcroft, 2019).

The correlations between scores from tests administered in both intervention and non-intervention groups are displayed in Table 4. The correlation matrix shows that all the variables correlate significantly with SV and PS at *p* < .001. All the correlations are also positive, which means that higher scores on one test are associated with higher scores on the other. What is more, all the correlations are moderate or strong. As anticipated SV and PS, which are intended to measure the same construct, correlate strongly which each other with *ρ* =.78, *p* <.001. The strong correlation between ORF and SV was also expected, but it is surprising that scores from these two tests correlate a bit more strongly (with *r* = .83, *p* <.001) than results from SV and PS. Another strong correlation was observed between SV and the part of TOWRE that intends to measure the decoding of words. This correlation of *r* = .75, *p* <.001 is stronger than the correlation between SV and part of TOWRE that assesseses the decoding of pseudowords (*r* = .57, *p* <.001). Similarly, the rate from Discrete Words correlates stronger with SV (*r* = .65, *p* <.001) than accuracy from the same test (*r* = .52, *p* <.001). As predicted SV correlates much weaker with NARA (*r* = .57, *p* <.001) than with ORF (*r* =.83, *p* <.001) However, it is surprising that PS correlates less strongly with NARA (*ρ* =.49, *p* <.001) than with ORF (*ρ* =.69, *p* <.001). Similarly to SV, PS correlates more strongly with TOWRE-words (*ρ* =.55, *p* <.001) than with TOWRE-pseudowords (*ρ* =.44, *p* <.001), and Spearman´s *ρ* of .59, *p* <.001 is higher for correlation between SP and rate from Discrete Words than for association between SP and accuracy (*ρ* =.46, *p* <.001). Generally, it seems that PS and SV have similar patterns of correlations with other tests, although PS correlates a bit weaker with other variables than SV. The interpretation of correlation patterns is, however, challenging because of the distribution characteristics of SV and PS that lead to the use of different correlation coefficients.

The next part of correlation analyses was conducted to establish divergent validity. PS and SV correlate weak and not significantly with RAVEN, TROG, and BPVS. There was observed no correlation between PS and RAVEN (*ρ* =0, *p* =.992), while the strongest positive association (but still not significant) was identified between BPVS and PS (*ρ* =.23, *p* = .144). TROG, BPVS, and RAVEN correlated negatively, very weakly, and not significantly with SV.

The last analyses indicated are conducted to examine the test-retest reliability (stability) of SV and PS. The scores from SV administered in Timepoint 1 and Timepoint 2 correlate positively and strongly with each other with r = .78, *p* <.001. This correlation coefficient can be considered as an acceptable level of reliability coefficient for a screener. On the other hand, the correlation coefficient between the two assessments of PS is much lower. Although the association with ρ =. 45***, *p* <.01 is considered as positive and moderate, it can be regarded as a low reliability coefficient for a screener.

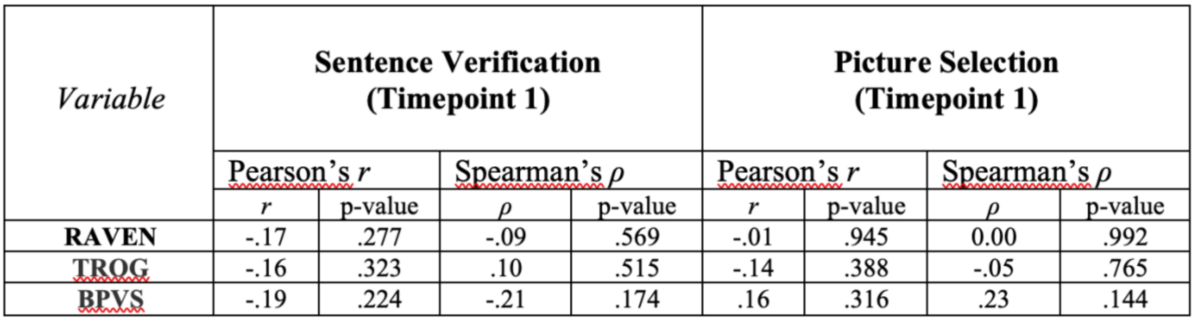
Table 4: Bivariate correlations (intervention & non-intervention group)



Variable	**Sentence Verification**		**Picture Selection**	
	Pearson's *r*	Spearman's *ρ*	Pearson's *r*	Spearman's *ρ*
NARA	.57	.55	.50	.49
TOWRE (words)	.75	.71	.58	.55
TOWRE (pseudowords)	.57	.54	.47	.44
ORF	.83	.83	.70	.69
Discrete words (accuracy)	.52	.49	.48	.46
Discrete words (rate)	.65	.67	.58	.59
Sentence Verification	----	----	.80	.78
Picture Selection	.80	.78	-----	----

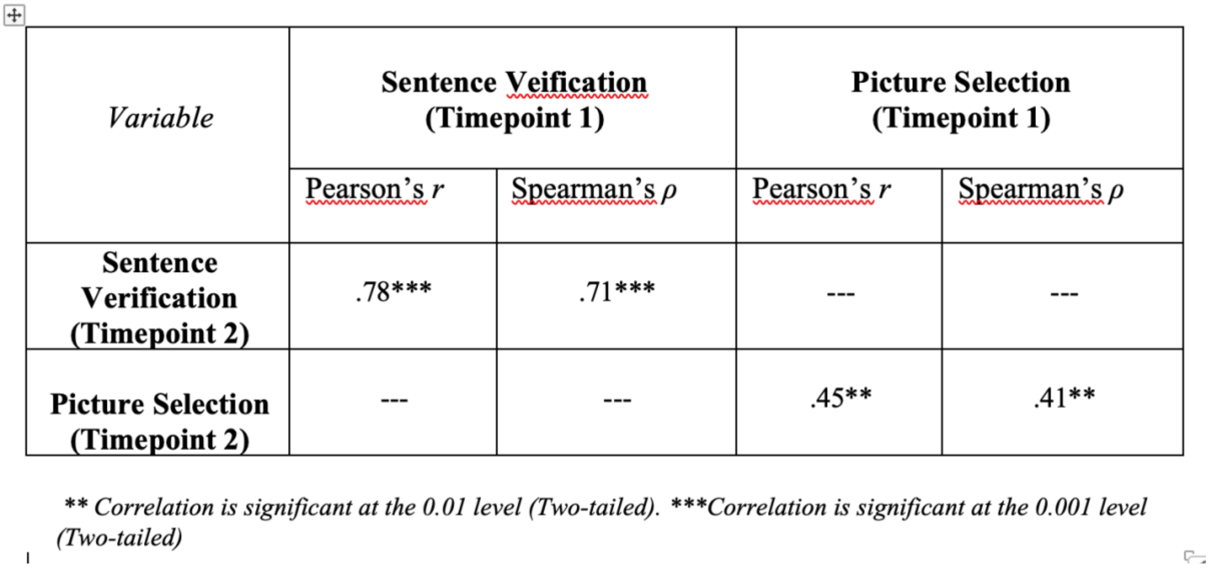
All the correlations in the table are significant at the .001level (Two-tailed).

Table 5: Bivariate correlations (intervention group): divergent validity



Variable	Sentence Verification (Timepoint 1)				Picture Selection (Timepoint 1)			
	Pearson's *r*		Spearman's *ρ*		Pearson's *r*		Spearman's *ρ*	
	r	p-value	*ρ*	p-value	*r*	p-value	*ρ*	p-value
RAVEN	-.17	.277	-.09	.569	-.01	.945	0.00	.992
TROG	-.16	.323	.10	.515	-.14	.388	-.05	.765
BPVS	-.19	.224	-.21	.174	.16	.316	.23	.144

Table 6: Bivariate correlations (intervention group): test-retest reliability



Variable	Sentence Veification (Timepoint 1)		Picture Selection (Timepoint 1)	
	Pearson's *r*	Spearman's *ρ*	Pearson's *r*	Spearman's *ρ*
Sentence Verification (Timepoint 2)	.78***	.71***	---	---
Picture Selection (Timepoint 2)	---	---	.45**	.41**

*** Correlation is significant at the 0.01 level (Two-tailed). ***Correlation is significant at the 0.001 level (Two-tailed)*

6. Discussion

The main research question of this thesis was whether the SV and PS show potential to be reliable and valid instruments that can be useful in research on silent reading efficiency and in screening processes in Norwegian schools. Since reliability is an important requirement for all assessment instruments, firstly, the test-retest method was used to establish the stability of SV and PS. The construct of silent reading efficiency combines notions of silent reading fluency (which includes accuracy and rate) and comprehension. Therefore, all those elements were addressed in statistical analysis with respect to concurrent validity as a second step of instruments validation. Additionally, measures of non-verbal intelligence and receptive vocabulary and grammar were taken into consideration during examining the divergent validity of the instruments.

The reliability analysis has shown that the stability of SV is rather low but acceptable for a screener, which should be used only for preliminary decisions. TOSREC, the instrument that inspired researchers from BetterReading, has shown similar test-retest reliability (between .78 and .92) in Grades 6 to 8 (Denton et al., 2011). In contrast, the test-retest reliability of PS seems to be too low to use that instrument as a screener, which indicates the need for further development of that measure to increase its stability. There are several reasons why the test-retest reliability was lower for PS than for SV. Murphy and Davishofer (1994) point out that tests that have fewer items tend to have lower reliability. The SV comprises 37 items, while PS has only 20. However, in time-limited tests the range of scores for both items might be more indicative. While SV has scores spread between 8 to 33 points, range of the scores for PS is from 1 to 13. The low number of correctly solved tasks may be partly caused because to short time provided to children, and partly because of a possible higher number of errors. The research assistants have also reported that children used the longer time to choose the right picture than it was anticipated during the construction of the test. The increasing of provided time could result in better test-retest reliability because the children could go through more items. Unfortunately, it would also put more demands on children´s motivation, attention, and stamina.

The second factor that may negatively influence the stability of PS is the difficulty of the higher difficulty level of the task in this test than in SV. First of all, each item from PS involves two steps: 1) reading the text 2) analyzing and choice of the picture. In contrast, SV requires only reading a sentence and choosing between two icons that look the same for all items. Secondly, understanding unrelated sentences does not require the use of higher language skills and makes SV an easier test. Thirdly, the pictures in PS have different numbers of details, which makes some items more difficult than others. An informal examination of error rates has shown that the average number of responses given for SV was 22.98 and the average number of errors was 1.07. At the same time, the average number of responses in PS was 8.49, while the average number of errors equaled 1.92. It suggests that pupils on average made mistakes almost five times more often during the administration of PS than under solving tasks from SV. It might indicate that PS was much more challenging for participants, and it gave them more room to guess the answers, despite the fact PS presented more answer options than SV. Further work on PS in. order to improve stability may include minimalizing the details on the picture to make the choice between them easier. However, it may be challenging to create pictures that reflect inferencing and integration without some necessary detailed, visually presented stimuli.

Other issues that could potentially impact the stability of PS and SV are technical problems during tests' administration, but there are no reasons to suspect that PS was more influenced by this than SV. Both tests had clear instructions that children could read and listen at the same time as well as two practice items that minimalized the chance of misunderstanding the tasks.

The low test-retest reliability of PS could unfortunately have an impact on validity measures. PS and SV are time-limited test, which means that children from the sample have completed different number of items. The number of items was determined by how fast the pupils have worked. Therefore, there was not possible to conduct internal consistency analyze for those instruments.

Both SV and PS intend to assess the same skill – efficient extracting of textual information during silent reading. However, the tests are constructed differently, which may lead to including other underlying skills in their measures. The sentence level of text processing in SV poses much less demands on inferencing, integration of information, and text monitoring than PS. Moreover, PS can require a larger capacity of working memory than SV, even though children have the possibility to reread the passage before solving the task in PS. Additionally, both the length of each item and the longer duration of the test administration may put greater demands on stamina and motivation in the case of PS. Finally, only PS requires the processing of visual stimuli that contain different numbers of details. Despite mentioned differences in construction, the strong and significant correlation between these tests and similar patterns of correlations with other measures may indicate that they assess the same or very similar construct.

Both SV and PS correlated strongly and significantly with ORF, which is in alliance with previous studies that showed associations between oral and silent reading fluency (Denton et al., 2011). Since the tests developed for BetterReading use sentences (SV) and passages (PS) as items, it is natural that they correlate more strongly with ORF than other fluency measures that use word lists or words presented individually. Moreover, SV in this study shows similar strength of the correlation with ORF as TOSREC in 5th Grade in study by Johnson and colleagues. While the correlation between TOSREC and ORF showed by was between .798 and .783 (Johnson et al., 2011), the correlation between SV and ORF examined in this thesis is .83. The sample from study by Johnson and colleagues consistent of American English-speaking children and therefore examined parameters of the test in context of deep (opaque) orthography, which is characterized by inconsistent relationship between pronunciation and written forms. In contrast, feature of transparent (shallow) orthographies is highly consistent relationship between spelling and pronunciation (Cain, 2010). Assessment of children in the present study was conducted in Norwegian language, which has semitransparent orthography. Similar strength of correlation between ORF and TOSREC in American English, and ORF and SV in Norwegian language seems to indicate that measures of reading comprehension effectivity may be useful for both orthographies. Moreover, since TOSREC showed good values in parameters that are important for a reliable screener - classification accuracy (90%), sensitivity (78%), specific (86%), and negative predictive value (98%) (Johnson et al., 2011) - it encourages further work on developing SV as a screener.

Because SV poses minimal demands on higher level text processing, researchers from BetterReading expected that SV would correlate more highly with ORF than with NARA, and these assumptions were confirmed by the results. However, it was surprising that also PS correlates more strongly with ORF than with NARA, although the component of comprehension should be bigger in PS than in SV. There are several possible explanations that do not exclude each other. Firstly, the low stability of PS could influence the results of correlation analyses connected to validity. Secondly, PS intends to capture differences in reading rate, and the role of the comprehension component is to ensure that children do not engage in “superficial” or “fake” reading. Therefore, the narrative text from PS is not very challenging for the most of pupils in 5th grade. In contrast, NARA is constructed to tap the individual differences in comprehension. The child who reads very slowly but answers comprehension questions correctly will probably get a better score on NARA than on PS. Thirdly, scores from NARA and PS may be affected by other skills that are not directly connected with reading comprehension. The open-ended questions from NARA put relatively big demands on the use of expressive language, while detailed pictures from PS require from participants visual processing. Although oral reading under the administration of NARA can support comprehension and staying on-task, the results of the test can be to greater extent affected by memory because children can read the text only once. Together these differences could make the comprehension component from PS much less like the comprehension accuracy measured by NARA than it was originally assumed.

Despite the correlation between PS and NARA being lower than expected, the fact that both instruments developed by BetterReading correlate moderately with NARA is congruent with previous studies that indicated an association between silent reading fluency and comprehension (Kim et al., 2015). Due to the not-experimental design of the study, the results of the statistical analysis do not allow drawing conclusions about the causality of this association or the direction of possible influence between comprehension and fluency. According to Psyridou et al. (2022), it is possible that, in the first years of primary school, it is fluency that impacts comprehension. This can be explained by verbal efficiency theory that claims that effective word reading, and good representations of lexical units are necessary for understanding of the text (Perfetti, 2007). The role of fluency in comprehension processes in early years is also underlined by the automaticity hypothesis that states that slow and inaccurate reading of words constrains cognitive resources, which cannot be delegated to higher level text processing (LaBerge & Samuels, 1974; Logan, 1997). On the other hand, it is also possible that knowledge of the text structure together with syntactic and semantic information from the text facilitate efficient reading. While some studies indicate that comprehension and fluency may be in simultaneous mutual interplay (Santos et al., 2020) other results indicate that the direction of causality changes during the development of reading skills (Psyridou et al., 2022). As stated before, our statistical analysis does not allow to draw conclusions about unidirectionality or bidirectionality of the association. The design of the study makes it also difficult (or impossible) to determine if the shift in the direction of impact has already occurred. However, according to Chall´s stages of reading development, average pupils from the 5th Grade should already have developed the necessary fluency skills that make reading for learning possible (Hierbert et al., 2012). If the average child from the sample performs on average level of the population, we can assume that decoding and reading rate do not constrain considerable comprehension accuracy of the average child from the sample. Unfortunately, there is no certainty regarding the relation of performances of participants from this study to performances of the whole population of Norwegian fifth graders.

The pattern of the correlations of SV and PS with TOWRE is in correspondence with the knowledge about reading development. The new tests correlate more strongly with subtest TOWRE-words than TOWRE-pseudowords. In the begging of literacy instruction children are highly dependent on phonological strategy when they read words. At the same time, there are greater individual differences in phonological skills among younger children, than in older pupils. It is possible that in 5th Grade most children have already developed the necessary skills to master phonological strategy, but there are still big differences when it comes to the size of sight word vocabulary. Moreover, children use more and more rarely phonological strategy to read familiar words (Price et al., 2016). Probably therefore accurate and fast word reading assessed by TOWRE-words correlate more strongly with scores from PS and SV than decoding pseudowords. Additionally, the results may be also interpreted as a confirmation of previous studies that showed smaller importance of phonological processing for silent reading than for oral reading (Hierbert et al., 2012; Price et al., 2012; Price et al., 2016; Juel & Holmes, 1981).

Similarly, correlations between PS and SV with two measures from Discrete Words (DW) are congruent with knowledge about the development of reading skills. Accurate word reading is a more constrained skill than the reading rate (Paris, 2004) and is acquired, at least to some level of proficiency, before children start to read words fluently (Mather & Wendling, 2012). It explains bigger variability between data points from the histogram of DW-rate, and smaller spreading of scores that are visible on the graph from DW-accuracy. It may indicate that DW-accuracy has ceiling effect because too many pupils read correctly almost all words from the test. It makes the test less useful for identifying individual differences in acquisition of reading skills among fifth graders (Murphy & Davidshoffer, 1994). The individual differences in rate are observable developmentally longer, and therefore there is no ceiling effect for DW-rate. It is also visible in bigger variance of datapoints on histogram of DW-rate. The lack of restricted variance allows for stronger correlation between the new tests and DW-rate.

The examination of concurrent validity has revealed patterns of correlations which indicate that scores from both instruments can reflect the silent reading fluency of 5th graders in relation to their comprehension of an accessible text. At the same time, there was no identified statistically significant correlation between instruments that intend to assess silent reading comprehension efficiency and BPVS or TROG. This may suggest that texts from PS and SV contain words and syntactic structures that were familiar to most of the students and did not impede neither their comprehension nor reading rate. Moreover, lack of significant correlations between the new tests and BPVS and TROG, may suggest that PS and SV do not differentiate between children with good and poor lower-language skills – vocabulary and grammar. The results of the third analyze connected to divergent validity has indicated that RAVEN did not correlate significantly with any of the new tests. That gives ground to believe that PS and SV do not tap skills connected to general non-verbal intelligence. That is especially important in the case of PS because of the previously stated concern that visual and spatial ability and attention to visual details could influence the scores from that test to the extant these are part of general non-verbal intelligence. In sum, the results of the examination of divergent validity indicates that reading comprehension efficiency measured by PS an SV is a distinct construct that is not directly associated with vocabulary, grammar and non-verbal intelligence.

After addressing reliability, concurrent validity, and divergent validity, it may be also relevant to take into consideration the ecological validity of PS and SV as potential screeners in Norwegian school. Although it is not possible to rule out measurement errors connected with technical difficulties, the app-based format of the instruments should not give additional challenges for Norwegian pupils who are using tablets in their free time and during the lessons. Many teachers use also digital intervention and assessment programs. Moreover, educational and career success in digital-global age may require in efficient reading on screen as well as on paper. But most importantly, the digital format gives the possibility for assessment in group-setting and automatic scoring, which saves teachers´ time that can be used for intervention. On the other hand, there are some mixed findings about the effect of reading on a screen that may result in compromised comprehension of longer texts (Singer et al., 2017). Additionally, the children that took part in the BetterReading project were assessed individually, and the presence of an examiner could help them stay on task. The group administration in the classroom may potentially result in some cases of "fake reading" and in consequence give an inaccurate picture of children´s reading skills. In this case, PS and SV would not be very useful instruments for screening of pupils in a group setting. Finally, pupils work more often with longer texts than unrelated sentences at school, which make the SV quite different from children´s everyday tasks. That is not a concern for PS which assesses silent reading comprehension efficiency in relation using longer connected passages.

7. Limitations

The features of the examined tests may pose some limitations to the overall construct validity. First of all, silent reading comprehension efficiency was defined as the rate of processing information from an accessible text during silent reading (Simonsen et al., 2022). However, the SV tap to a very limited extent higher language skills that readers use in a natural situation. Single, unrelated sentences do not require ongoing comprehension monitoring, the use of knowledge about different text structures, or the integration of information from different parts of a text. Similarly, only local inferences can be made. Although the readers still need to verify the content of sentences with their general knowledge, it seems that SV does not measure all the components that the construct of SRCE includes.

On the other hand, scores from PS might be affected by skills that are not included in the construct of SRCE. The results of statistical analysis have shown that PS did not correlate significantly with RAVEN which intends to measure among others visual attention. However, RAVEN gives a compound measure of non-verbal intelligence that includes also logical thinking, working memory, and spatial and categorization ability. Therefore, the lack of significant correlation between PS and RAVEN does not exclude the possibility that scores on PS might be impacted by skills of visual processing and attention to detailed, visually presented stimuli that are not included in the construct of SRCE.

There are also limitations concerning the external validity of the results, which do not permit generalizing findings to all 5th graders in Norway. Firstly, the sample included only children that have attended Norwegian school from the first grade. Moreover, pupils from the intervention group had to read fluently enough to be able to join – and benefit from – the intervention program. This demand was secured by preliminary testing. Secondly, the participants were self-selected and not random, which is also considered to be a threat to external validity (Lund, 2002). Thirdly, the choice of the schools that were engaged in the project was not random but motivated by researchers' familiarity with the personnel. Therefore, all the schools were situated in Oslo area and constitute a convenience sample that gives limited possibilities for generalization to other parts of Norway.

Other limitations concern the statistical validity of analysis involved in the examination of test-retest reliability and internal consistency analyses of ORF, RAVEN, BPVS, and TROG. Because of constrained time resources, only a part of the children from the sample was assessed with these tests, and that could influence the strength and significance of the findings.

8. Conclusion

Efficient silent reading of an accessible text may demand various underlying skills that depend on the level of text processing and a required task. Despite differences in the construction of SV and PS, the tests correlate strongly with each other and show the same pattern of associations with measures of comprehension accuracy, oral reading fluency and its components. Moreover, they do not show statistically significant associations with instruments that tap skills which are not directly embedded in the construct of silent reading comprehension efficiency. That gives ground to state that they indeed assess the same or very similar construct. However, the features of the tasks in both tests may poses some threats to construct validity. SV using only unrelated sentences as a unit, includes in very small extant higher language skills that ply important role for effective silent reading of the text and uses sentences. In contrast, scores in PS may be influenced by processing of visually presented stimuli which contains a lot of details, and that skill is not included in the concept of SRCE.

Although fast and accurate extracting of textual information together with its application in practical tasks is an important skill in modern society, it seems that presently there is no assessment instrument that could help to identify inefficient readers in Norwegian schools. Because of higher reliability and stronger correlations with relevant measures, SV showed bigger potential as a screener and research tool than PS. Despite its limitations concerning ecological validity and construct validity, thanks to possibility of group administration and automatic scoring, the test could be a more economical and time-saving alternative to ORF, and a supplement to *Ordkjedetesten*.

8.1 Implications

The lack of reliable and effective screeners which could help to identify late emerging difficulties among Norwegian pupils together with the favorable results of this study regarding SV gives several implications for further work on developing that instrument. Firstly, it may be important to check how SV correlates with compulsory national tests (Nasjonale prøver) in reading that are administered in 5th, 8th, and 9th grade to strengthen evidence of concurrent validity and determine predictive validity. Establishing accuracy in identifying pupils that are at risk of poor performance in future assessment requires, however, a longitudinal design. Secondly, using SV as a screener also requires the development of norms for each age group for which the instrument is intended to use. In this case, the future study design should secure that the sample is relatively large and representative for the Norwegian pupil population. It should be examined if SV has high enough classification accuracy, sensitivity, specificity, positive and negative predictive value to be an effective and reliable screener. Thirdly, it is important to strengthen evidence of the ecological validity of the test by checking how well scores from SV reflect pupils' reading skills when the test is administrated in a classroom setting. Finally, it could be interesting to compare SV and *Ordkjedetesten* to examine possible differences in skills that these two tests tap. Although both include rate, and a decision component, they use a different level of text processing and have different administration formats (digital vs. paper-and-pencil). The results could inform further research on silent reading rate and silent reading comprehension efficiency.

As regards PS, it is important to notice that a small sample size could influence the strength and significance of the correlations included in the analysis of stability. Moreover, the test-retest reliability was computed with data gathered from subsample of intervention children, who were previously identified as struggling with reading on an age-appropriate level. That may have resulted with small range of scores attained by these children and weaker correlations. It can be worthwhile to examine if the test could show better reliability in studies that use a larger and more representative sample. Moreover, it might be favorable to increase homogeneity between the items' difficulty level and ensure easier choice between the pictures by limitation of details that their include. A more detailed analysis of test features and procedures concerning test administration could be made to identify other factors that may contribute to low test-retest reliability.

References:

Aaron, P. G., Joshi, M. & Williams, K. A. (1999) Not all reading disabilities are alike. *Journal of learning disabilities, 32*(2), 120-137 https://journals-sagepub-com.ezproxy.uio.no/doi/epdf/10.1177/002221949903200203

Armbruster, B. B., Lehr, F., & Osborn, J. (2001) *Put reading first: The research building blocks for teaching children to read*. National Institute for Literacy. https://files.eric.ed.gov/fulltext/ED458536.pdf

Arnesen, A., Braeken, J., Ogden, T., Melby-Lervåg, M. (2018) Assessing Children´s Social Functioning and Redaing Proficiency: A Systematic Review of the Quality of Educational Assessment Instruments Used in Norwegian Elementary Schools. *Scandinavian journal of educational research, 63*(3), s.465-490

Arnesen, A., Braeken, J., Baker, S., Meek-Hansen, W., Ogden, T. & Melby-Lervåg, M., (2017) Growth in Oral Reading Fluency in a Semitransparent Orthography: Concurrent and Predictive Relations with Reading Proficiency in Norwegian, Grades 2-5. *Reading Research Quarterly*, *52*(2), 177–201. https://doi.org/10.1002/rrq.159

Bar-Kochva, I. (2013) What are the underlying skills of silent reading acquisition? A developmental study from kindergarden to the 2nd grade. *Reading & writing, 26*(9), 1417-1436 https://link-springer-com.ezproxy.uio.no/content/pdf/10.1007/s11145-012-9414-3.pdf

Carlsten, C. T. (2016). *Carlstenprøvene*. Cappelen Damm.

Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and measurement* (9th ed.). Boston: McGraw-Hill.

Cain, K., & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology, 76*(4), 697–708. https://doi.org/10.1348/000709905x69807

Cain, K. (2010) *Reading development and difficulties*. BPS Blackwell

Cao, Y. & Kim, Y.-S. G.(2021) Is retell a valid measure of reading comprehension? *Educational research review*, 32, https://www-sciencedirect-com.ezproxy.uio.no/science/article/pii/S1747938X20308356

Denton, C. A., Barth, A. E., Fletcher, J. M., Wexler, J., Vaughn, S., Cirino, P. T. Romain, M., & Francis, D. J. (2011). The Relations Among Oral and Silent Reading Fluency and Comprehension in Middle School: Implications for Identification and Instruction of Students with Reading Difficulties. *Scientific studies of reading*, *15*(2), 109-135

Ecalle, J. Bouchafa, H., Potocki, A. & Magnan, A. (2013) Comprehension of written sentences as a core component of children´s reading comprehension. *Journal of research in reading, 36*(2), 117-131 https://onlinelibrary-wiley-com.ezproxy.uio.no/doi/pdfdirect/10.1111/j.1467-9817.2011.01491.x

Elliot, J. G. & Grigorenko, E. L. (2014) *The Dyslexia Debate.* Cambridge University Press

Fletcher, J. M., Lyon, G. R., Fuchs, L. S. & Barnes, M. A. (2019) *Learning disabilities. From Identification to Intervention* (2st ed.). The Guilford Press.

Good, R. H., Kaminski, R.A., & Dill, S. (2002). DIBELS oral reading fluency and retell fluency. In R. H. Good & R.A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th ed., pp. 30– 38). Eugene, OR: Institute for the Development of Education Achievement.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and special education, 7*(1), 6-10. https://journals-sagepub-com.ezproxy.uio.no/doi/epdf/10.1177/074193258600700104

Griffith, L. W. & Rasinski, T. (2004). A focus on fluency: how one teacher incorporated fluency with her reading curriculum. *The Reading teacher, 58*(2), 126-137 https://go-gale-com.ezproxy.uio.no/ps/i.do?p=AONE&u=oslo&id=GALE%7CA123677410&v=2.1&it=r

Helland-Riise, F. & Martinussen, M. (2017) *Måleegenskaper ved de norske versjonene av Ravens matriser [Standard Progressive Matrices (SPM) / Coloured Progressive Matrices (CPM)].* PsykTestBarn, 2:2 https://r-bup.brage.unit.no/r-bup-xmlui/handle/11250/2728772

Hiebert, E. H., Samuels, S. J. & Rasinski,T. (2012). Comprehension-Based Silent Reading Rates: What do we know? What do we need to know? *Literacy Research and Instruction,* 110-124 https://www-tandfonline-com.ezproxy.uio.no/doi/full/10.1080/19388071.2010.531887

Hiebert, E. H., Wilson, K. M., & Trainin, G. (2010). Are students really reading in independent read- ing contexts? An examination of comprehension-based silent reading rate. In E. H. Hiebert & D. R. Reutzel (Eds.), *Revisiting silent reading: New directions for teachers and researchers,* 151–167. Newark, DE: International Reading Association.

Hierbert, E. H. & Daniel, M. (2019). Comprehension and rate during silent reading: Why do some students do poorly? *Reading & writing, 32*(7), 1795-1818 https://link-springer-com.ezproxy.uio.no/article/10.1007/s11145-018-9917-7

Hintze, J. M., Callahan, J. E., Matthews, W. J., Williams, S. A. S & Tobin, K. G. (2002) Oral Reading Fluency and Prediction of Reading Comprehension in African American and Caucasian Elementary School Children. *School psychology review, 31*(4), 540-553 https://www-tandfonline-com.ezproxy.uio.no/doi/pdf/10.1080/02796015.2002.12086173?needAccess=true&

Hogan, T. P., Bridges, M. S., Justice, L. M., Cain, K. (2011). Increasing Higher Level Language Skills to Improve Reading Comprehension. *Focus on Exceptional Children, 44* (3), 1-20 https://go-gale-com.ezproxy.uio.no/ps/i.do?p=AONE&u=oslo&id=GALE|A277534867&v=2.1&it=r

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading & Writing, 2*(2), 127–160. https://link-springer-com.ezproxy.uio.no/content/pdf/10.1007/BF00401799.pdf

Høien, T. & Tønnesen, G. (2008). *Instruksjonshefte til ordkjedetesten* (5th ed.). Logometrica AS

Jenkins, J. R., Fuchs, L. S., van Den Broek, P., Espin, Ch. & Deno, S.L. (2003) Accuracy and Fluency in List and Context Reading of Skilled and RD Groups: Absolute and Relative Performance Levels. *Learning Disabilities Research & Practice*. *18*(4), 237-245 https://onlinelibrary-wiley-com.ezproxy.uio.no/doi/abs/10.1111/1540-5826.00078

Jenkins, J. R. Hudson, R. F., & Johnson, E. S. (2007) Screening for at-risk readers in response to intervention framework. *School Psychology Review, 36*, 582-600. https://go-gale-com.ezproxy.uio.no/ps/i.do?p=AONE&u=oslo&id=GALE|A173466636&v=2.1&it=r

John, O. P. & Benet-Martinez, V. (2014) Measurment: Reliability, Construct Validation, and Scale Construction. In H. T. Reis & Ch. M. Judd (Eds.) *Handbook of research methods in Social and Personality Psychology* (pp. 473-503). Cambridge University Press. https://www-cambridge-org.ezproxy.uio.no/core/books/handbook-of-research-methods-in-social-and-personality-psychology/measurement/EB7A7EC2664B491707897B6C399F9477

Johnson, E. S., Pool, J. L. & Carter, D. R. (2011). Validity Evidence for the Test of Silent Reading Efficiency and Comprehension (TOSREC). *Assessment for effective intervention, 37*(1), 50-57 https://journals-sagepub-com.ezproxy.uio.no/doi/full/10.1177/1534508411395556

Juel, C. & Holmes, B. (1981) Oral and Silent Reading of Sentences. *Reading Research Quarterly, 16*(4), 545-568 https://www-jstor-org.ezproxy.uio.no/stable/pdf/747315.pdf?refreqid=excelsior%3Af61944bf60708cbeccdf6e43355ba373&ab_segments=&origin=&initiator=&acceptTC=1

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology, 102*(3), 652–667. https://psycnet-apa-org.ezproxy.uio.no/fulltext/2010-15712-010.html

Kintsch, W. (1988) The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review, 95*(2), 163-182 https://psycnet-apa-org.ezproxy.uio.no/fulltext/1988-28529-001.html

Kintsch, W. & Rawson, K. A. (2005) Comprehension. In M. J. Snowling & C. Humle (Eds.) *The science of reading: a handbook*. Blackwell Publishing

Kleven, T. A. (2002). Ikke-eksperimentelle design. In T. Lund (Ed.), *Innføring i forskningsmetodologi*. Bergen: Fagbokforlaget Vigmostad & Bjørke AS

Kuhn, M. R., Schwanenflugel, P. J. &Meisinger, E. B. (2010) Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(5), 232-253 https://ila-onlinelibrary-wiley-com.ezproxy.uio.no/doi/pdfdirect/10.1598/RRQ.45.2.4

Kunnskapsdepartamentet (2017) Meld. St. 21 (2016-2017). *Lærelyst – tidlig innsats og kvalitet i skolen*. https://www.regjeringen.no/no/dokumenter/meld.-st.-21-20162017/id2544344/

LaBerge, D. & Samuels, S. J. (1974) Toward a theory of automatic information processing in reading. *Cognitive psychology, 6*(2), 293-323 https://www-sciencedirect-com.ezproxy.uio.no/science/article/pii/0010028574900152

Lauterman, T. & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in human behavior, 35*, 455-463 https://libkey.io/libraries/269/articles/49975867/full-text-file

Logan, G. D. (1997) Automaticity and reading: perspectives from the instance theory of automatization. *Reading & writing quarterly, 13*(2), 123-146 https://www-tandfonline-com.ezproxy.uio.no/doi/abs/10.1080/1057356970130203

Lund, T. (2002). Metodologiske prinsipper og referanserammer. In T. Lund (Ed.), *Innføring i forskningsmetodologi*. Bergen: Fagbokforlaget.

Lyster, S. A. H. & Horn, E. (2009) *Trog-2. Norsk versjon. Test for Reception of Grammar – Version 2. Manual*. Pearson Assessment

Lyster, S.-A.H., E. Horn & Rygvold, A. L. (2010). Ordforråd og ordforrådsutvikling hos norske barn og unge. Resultater fra en utprøving av British Picture Vocabulary Scale, Second Edition (BPVS II). *Spesialpedagogikk 09*, 35-43.

Mather, N. & Wendling, B. J. (2012) *Essentials of Dyslexia. Assessment and Intervention*. Wiley Blackwell

Mangen, A., Walgermo, B. R. & Brønnick, K. (2013) Reading linear text on paper versus computer screen: Effects on reading comprehension. *International journal of educational* *research, 58,* 61-68 https://www-sciencedirect-com.ezproxy.uio.no/science/article/pii/S0883035512001127

NESH (2016) Forskningsetiske retningslinjer for samfunnsvitenskaplige, humanoria, juss og teologi. In. Oslo: Oktan Forlag.

National Center on Response to Intervention (2010) *Essential components of RTI – A closer look at response to intervention.* U.S.Office of Special Education Programs https://files.eric.ed.gov/fulltext/ED526858.pdf

National Reding Panel (2000). Report of the national reading panel: Teaching children to read, Reports of the subgroups. NIH Pub. 00-4754. U.S. Department of Health and Human Services.

Navarro, D. J. & Foxcroft, D. R. (2019) *Leraning statistics with Jamovi: a tutorial for psychology students and other beginners* (5th ed.). The Jamovi Project. Doi: 10.24384/hgc3-7p15

Neale, M. D. (1958). *Neale Analysis of Reading Ability*. MacMillan.

Norton, E. S. & Wolf, M. (2012) Rapid Automatized Naming (RAN) and Reading Fluency: Implications for Understanding and Treatment of Reading Disabilities. *Annual review of psychology, 63*(1), 427-452 https://www-annualreviews-org.ezproxy.uio.no/doi/10.1146/annurev-psych-120710-100431

Oakhill, J., Cain, K. & Elbro, C. (2006) Reading Comprehension and Reading Comprehension Difficulties. In D. A. Kilpatrick, R. M. Joshi & R. K. Wagner (Eds.) *Reading Development and Difficulties*. Springer

O´Brien, B. A, Wallot, S., Haussmann, A. & Kloos, H. (2014) Using Complexicity Metrics to Assess Silent Reading Fluency: A Cross-Sectional Study Comparing Oral and Silent Reading. *Scientific Studies of Raeding*, 18(4), s.235-254 https://www-tandfonline-com.ezproxy.uio.no/doi/full/10.1080/10888438.2013.862248

Perfetti, Ch. (2007) Reading Ability: Lexical Quality to Comprehension. *Scientific studies of reading, 11*(4), 357-383 https://www-tandfonline-com.ezproxy.uio.no/doi/pdf/10.1080/10888430701530730?needAccess=true

Paris, S. G. (2004) Reinterpreting the development of reading skills. *Reading research quarterly, 40*(2), 184-202 https://ila-onlinelibrary-wiley-com.ezproxy.uio.no/doi/pdfdirect/10.1598/RRQ.40.2.3

Price, K. W., Meisinger, E. B., Louwerse, M. M. & D´Mello, S. (2012). Silent reading fluency using underlining: Evidence for an alternative method of assessment. *Psychology in the schools*, *49*(6), 606-618 https://onlinelibrary-wiley-com.ezproxy.uio.no/doi/full/10.1002/pits.21613

Price, K. W., Meisinger, E. B., Louwerse, M. M. & D´Mello, S. K. (2016) The Contributions of Oral and Silent Reading Fluency to Reading Comprehension. *Reading Psychology,* 37(2), s. 167 -201 https://www-tandfonline-com.ezproxy.uio.no/doi/full/10.1080/02702711.2015.1025118

Protopapas, A., Simos, P. G., Sideridis, G. D., & Mouzaki, A. (2012). The components of the simple view of reading: A confirmatory factor analysis. *Reading Psychology, 33*(3), 217–240. https://www-tandfonline-com.ezproxy.uio.no/doi/pdf/10.1080/02702711.2010.507626?needAccess=true

Psyridou, M., Tolvanen, A.; Niemi, P., Lerkkanen, M.-K., Poikkeus; A.-M., Torppa, M. (2022). Development of silent reading fluency and reading comprehension across grades 1 to 9: unidirectional or bidirectional effects between the two skills? *Reading and writing,* https://link-springer-com.ezproxy.uio.no/article/10.1007/s11145-022-10371-6

Punch, K. F. & Oancea, A. (2014) *Introduction to Research Methods in Education*. Sage.

Rasinski, T., Samuels, S. J., Hierbert, E., Petscher, Y., Feller, K. (2011). The relationship between a silent reading fluency instructional protocol on student´s reading comprehension and achievement in an urban school setting. *Reading psychology, 32*(1), 75-97 https://www-tandfonline-com.ezproxy.uio.no/doi/full/10.1080/02702710903346873

Rønberg, L. F. & Petersen, D. K. (2016) It matters whether reading comprehension is conceptualized as rate or accuracy: reading comprehension rate versus accuracy. *Journal of research in reading, 39*(2), 209-228 https://onlinelibrary-wiley-com.ezproxy.uio.no/doi/pdfdirect/10.1111/1467-9817.12047

Santos, S., Cadime, I., Viana, F. L. & Ribeiro, I. (2020) Cross-Lagged Relations Among Linguistic Skills in European Portugese: A longitudinal Study. *Reading research quarterly, 55*(2), 177-192 https://ila-onlinelibrary-wiley-com.ezproxy.uio.no/doi/pdfdirect/10.1002/rrq.261

Schwanenflugel, P. J. & Knapp, N. F. (2016) *The psychology of reading: theory and applications.* The Guilford Press.

Schwanenflugel, P. J. & Kuhn, M. R. (2016) Reading Fluency. In P. Afflerbach (Ed.), *Handbook of individual differences in reading. Reader, Text, and Context* (pp.107-115). Routledge

Simonsen, K., Altani, A., Zelihic, Dz., Ziaka, L., Braze, D. & Protopapas, A. (2022, November, 17-20). *Development of two tests of reading comprehension efficiency* [poster]. Psychonomic Society, 63rd Annual Meeting, Boston, Massachusetts, United States 10.13140/RG.2.2.10736.74242

Singer, L. M. & Alexander, P. (2017). Reading on Paper and Digitally: What the Past Decades of Empirical Research Reveal. *Review of educational research*, *87*(6), 1007-1041 https://journals-sagepub-com.ezproxy.uio.no/doi/full/10.3102/0034654317722961

Snow, C. E & Vaughn, S. (2018) Simple and not-so-simple views of reading. *Remedial and special education, 39*(5), 313-316 https://journals-sagepub-com.ezproxy.uio.no/doi/epub/10.1177/0741932518770288

Spichtig, A. N., Hiebert, E. H., Vorstius, Ch., Pascoe, J. P., Pearson, P. & Radach, R. (2016). The Decline of Comprehension-Based Silent Reading Efficiency in the United States: A Comparision of Current Data With Performance in 1960. *Reading Research Quarterly, 51*(2), 239-259 https://ila.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/rrq.137

Spichtig, A. N., Pascoe, J. P, Gehsmann, K. M., Gu, F. & Ferrara, J. D. (2022) The Interaction of Silent Reading Rate, Academic Vocabulary, and Comprehension Among Students in Grade 2-12. *Reading research quarterly, 57*(3), 1003-1019 https://ila-onlinelibrary-wiley-com.ezproxy.uio.no/doi/pdfdirect/10.1002/rrq.457

Stanovich, K. E. (1986) Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360-407 https://www-jstor-org.ezproxy.uio.no/stable/747612?sid=primo#metadata_info_tab_contents

Tarar, J. M., Mesinger, E.B., & Dickens, R. H. (2015) Test Review: Test of Word Reading Efficiency -Second Edition (TOWRE-2) by Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. *Canadian Journal of School Psychology. 30*(4) https://journals-sagepub-com.ezproxy.uio.no/doi/full/10.1177/0829573515594334

Taylor, S. E. (1965). Eye Movements in Reading: Facts and Fallacies. *American research journal*, *2*(4), 187-202 https://www-jstor-org.ezproxy.uio.no/stable/1161646?sid=primo

The Jamovi Project. (2022). Jamovi (Version 2.3.21). Retrieved from https://www.jamovi.org/download.html

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). *Test of Word Reading Efficiency–Second Edition*. Austin, TX: Pro-Ed.

Vellutiono, F. R., Tunmer, W. E., Jaccard, J. J & Chen, R. (2007) Components of Reading Ability: Multivariate Evidence for Convergent Skills Model of Reading Development. *Scientific studies of reading*, *11*(1), 3-32 https://www-tandfonline-com.ezproxy.uio.no/doi/pdf/10.1080/10888430709336632?needAccess=true

van den Boer, M., van Bergen, E., de Jong, P. F. (2014). Underlying skills of oral and silent reading. *Journal of Experimental Child Psychology*, 128, 138-151

van den Broek, P., Espin, C. A. & Burns, M. K. (2012) Connecting cognitive theory and assessment: measuring individual differences in reading comprehension. School psychology review, 41(3), 315-325 https://www-tandfonline-com.ezproxy.uio.no/doi/pdf/10.1080/02796015.2012.12087512?needAccess=true&

van den Broek, P., Mouw, J. M. & Kraal, A. (2016) Individual Differences in Reading Comprehension. In P. Afflerbach (Ed.), *Handbook of individual differences in reading. Reader, Text, and Context* (pp.107-115). Routledge

Walgermo, B. R., Uppstad, P. H., Lundetræ, K., Tønnessen, F. E. & Solheim, O. J. (2018) Kartleggingsprøver i lesing – tid for nytenkning? Acta didactica Norge, 12 (4) file:///Users/zuzannasolska/Downloads/torgeich,+121218.walgermo.mfl.%20(1).pdf

Webman-Shafran, R. (2018) Implicit prosody and parsing in silent reading. *Journal of research in reading, 41*(3), 546-563 https://onlinelibrary-wiley-com.ezproxy.uio.no/doi/pdfdirect/10.1111/1467-9817.12124

Wexler, J., Vaughn, S., Edmonds, M., & Reutebuch, C. K. (2008). A synthesis of fluency interventions for secondary struggling readers. *Reading and Writing*, *21*, 317–347. https://link-springer-com.ezproxy.uio.no/content/pdf/10.1007/s11145-007-9085-7.pdf

Wissinger, D. R., Truckenmiller, A. J., Konek, A. E. & Ciullo, S. (2023). The Validity of Two Tests of Silent Reading Fluency: A Meta-Analytic Review. *Reading & writing Quarterly*, Ahead-of-print, 1-17 https://www-tandfonline-com.ezproxy.uio.no/doi/full/10.1080/10573569.2023.2175340

Appendix 1: Normal Q-Q-plots of the Variables

Figure 19: Normal Q-Q-Plot of Sentence Verification (both groups combined)

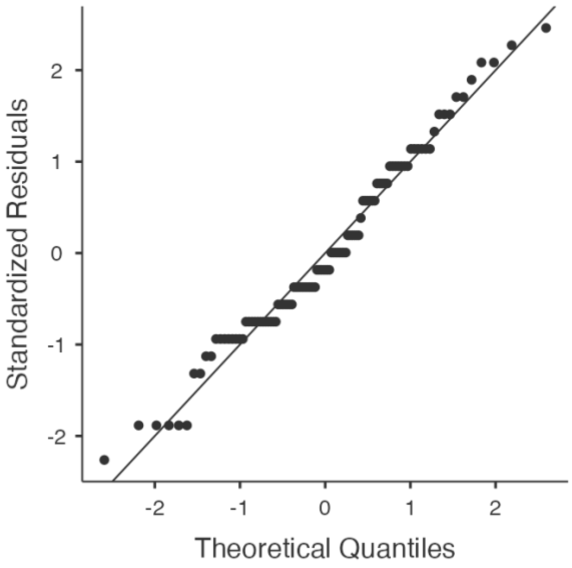


Figure 20: Normal Q-Q-Plot of Picture Selection (both groups combined)

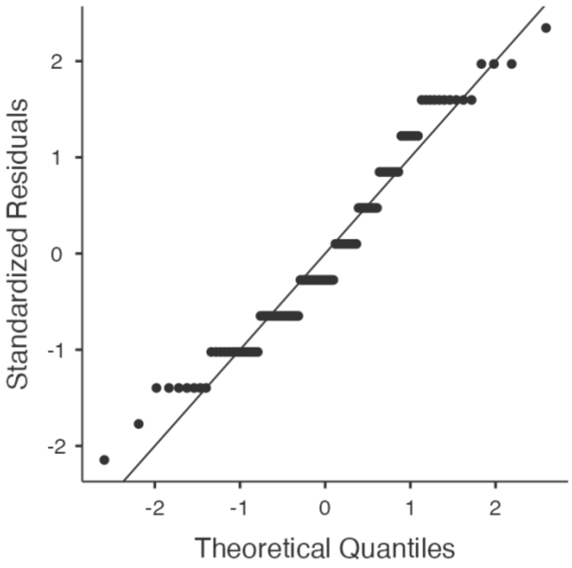


Figure 21: Normal Q-Q-Plot of NARA (both groups combined)

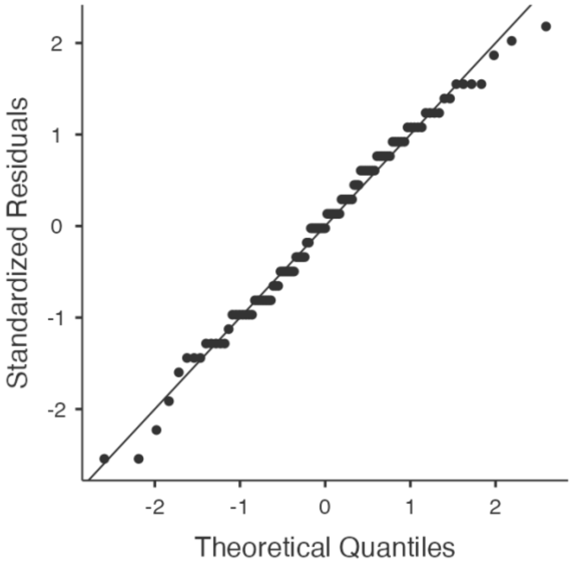


Figure 22: Normal Q-Q-Plot of TOWRE – Words (both groups combined)

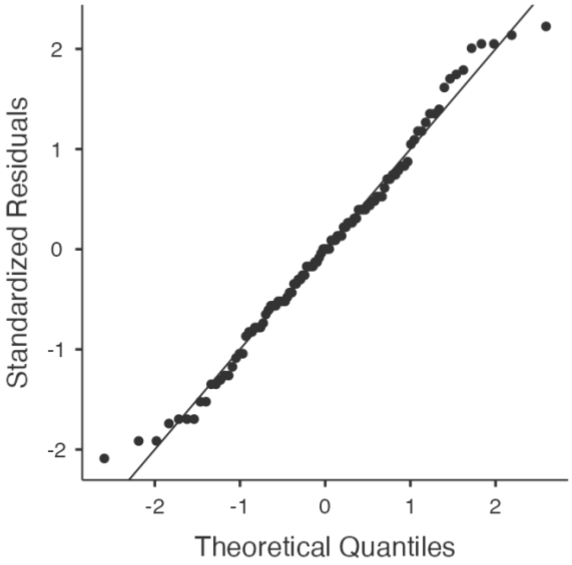


Figure 23: Normal Q-Q-Plot of TOWRE – pseudowords (both groups combined)

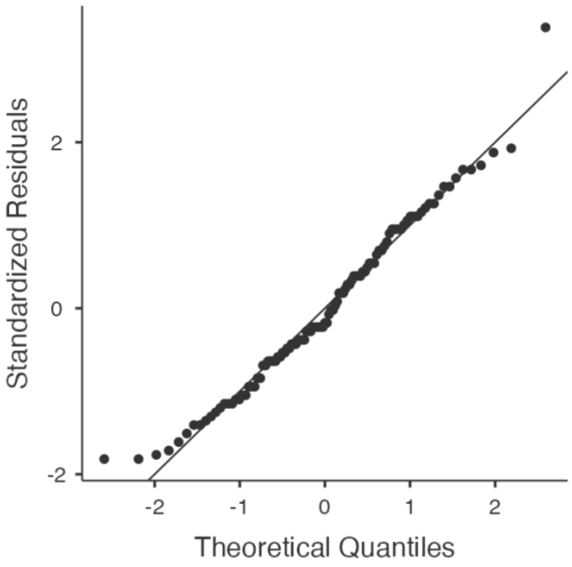


Figure 24: Normal Q-Q-Plot of ORF (both groups combined)

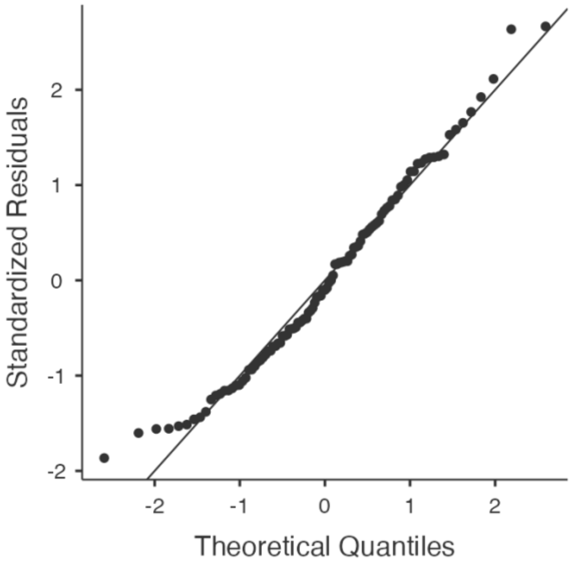


Figure 25: Normal Q-Q-Plot of Discrete Words – Accuracy (both groups combined)

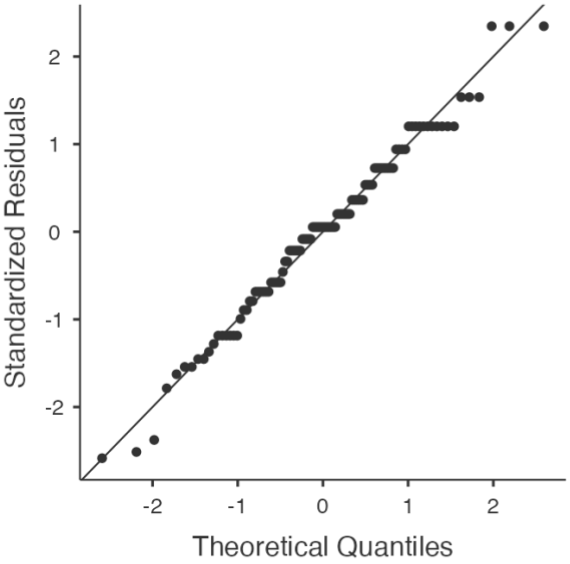


Figure 26: Normal Q-Q-Plot of Discrete Words – Rate (both groups combined)

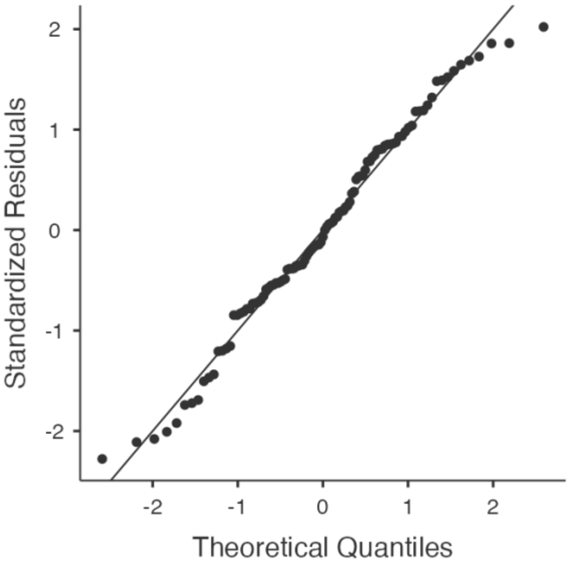


Figure 27: Normal Q-Q-Plot of RAVEN (intervention group)

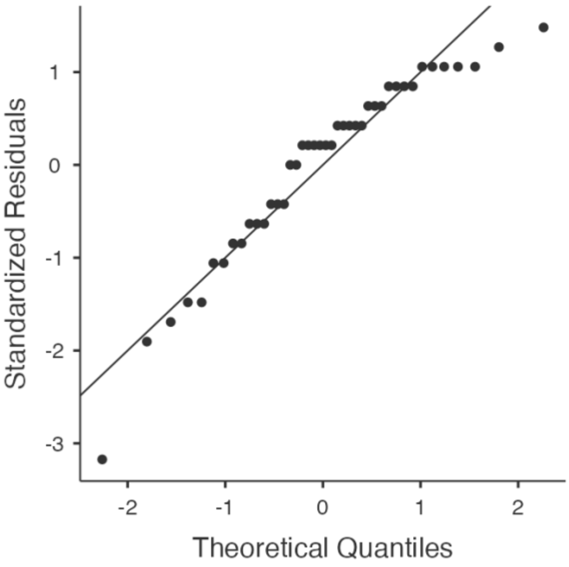


Figure 28: Normal Q-Q-Plot of TROG (intervention group)

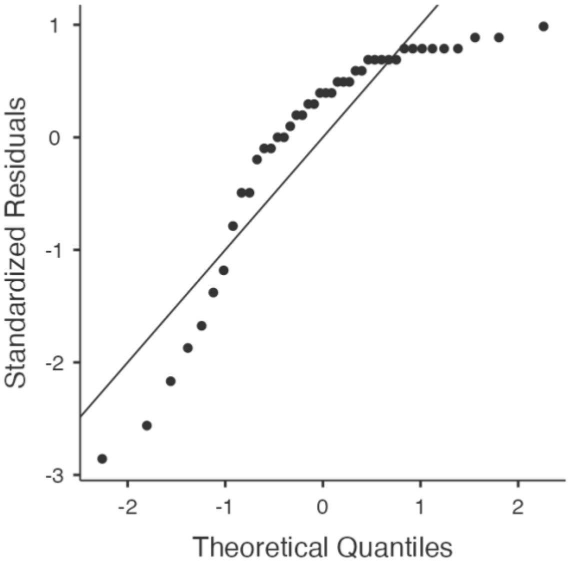


Figure 29: Normal Q-Q-Plot of BPVS (intervention group)

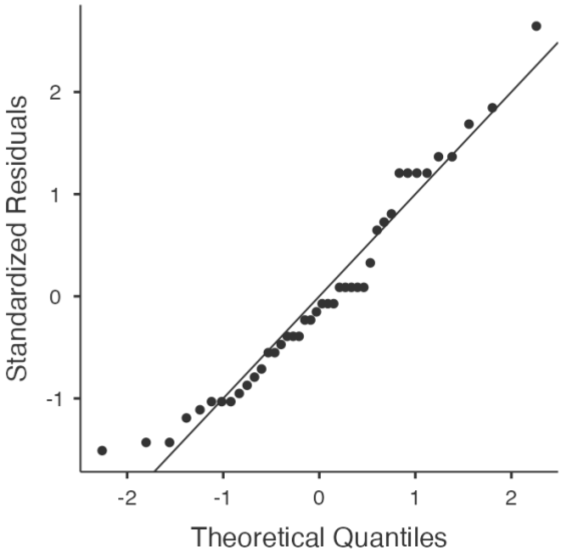


Figure 30: Normal Q-Q-Plot of Sentence Verification – timepoint 1 (intervention group)

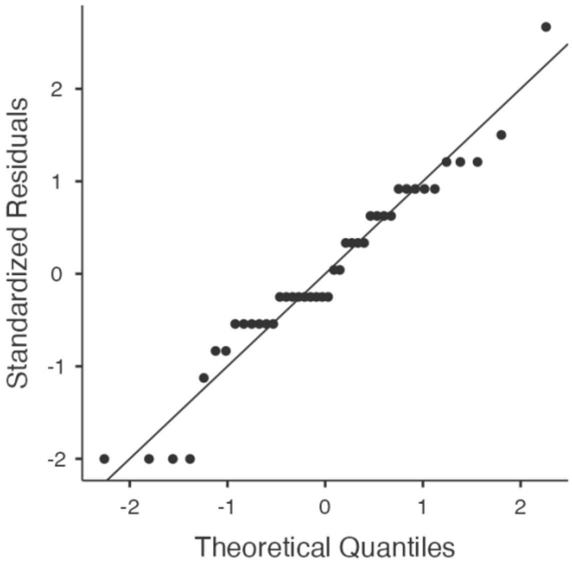


Figure 31: Normal Q-Q-Plot of Sentence Verification – timepoint 2 (intervention group)

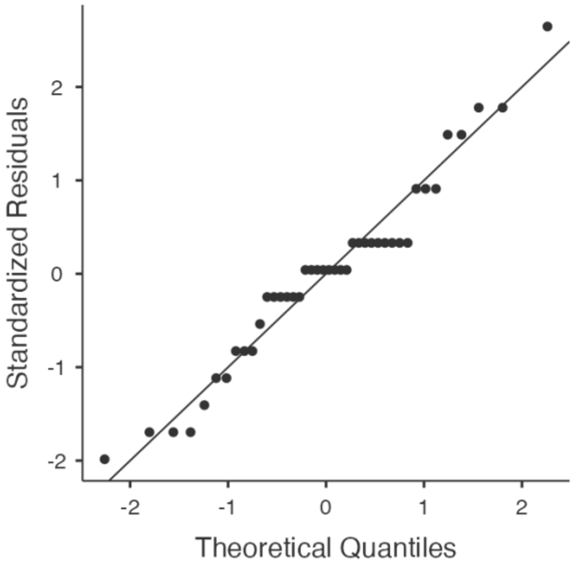


Figure 32: Normal Q-Q-Plot of Picture Selection - timepoint 1 (intervention group)

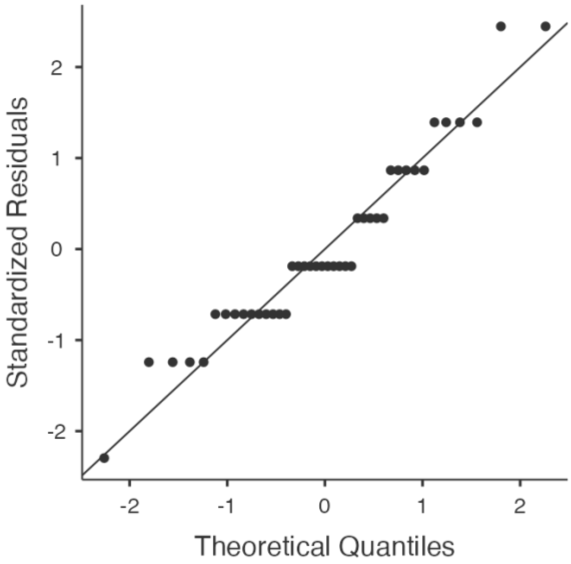
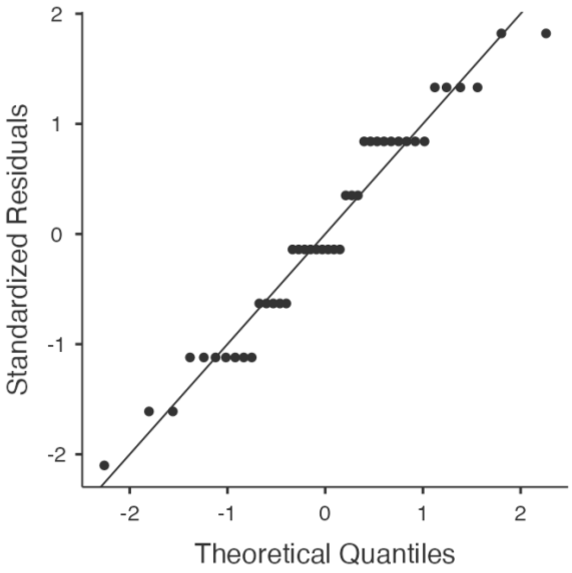


Figure 33: Normal Q-Q-Plot of Picture Selection - timepoint 2 (intervention group)



Appendix 2: Scatterplots

2.1 SCATTERPLOTS – BOTH GROUPS COMBINED

Figure 34: Scatterplot of Sentence Verification & Picture Selection

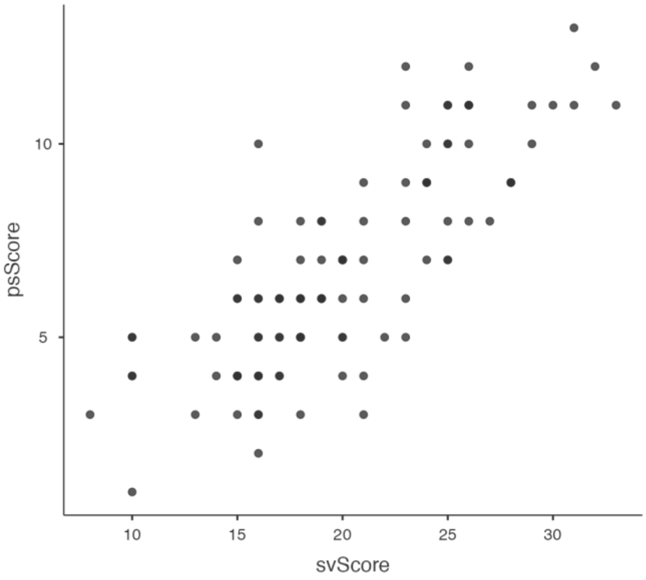


Figure 35: Scatterplot of Sentence Verification & NARA

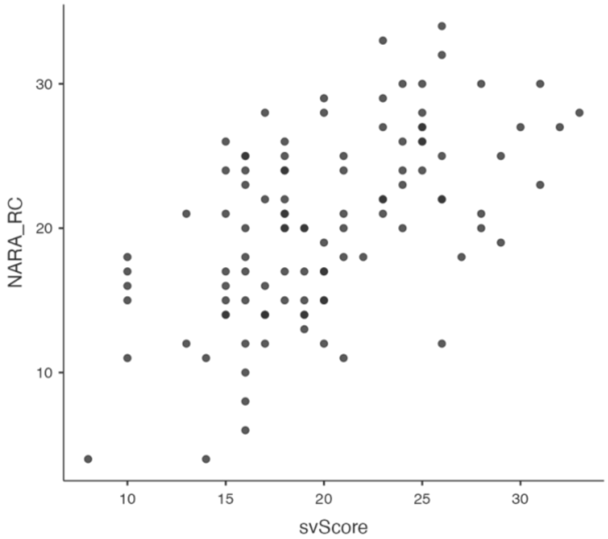


Figure 36: Scatterplot of Sentence Verification & TOWRE - words

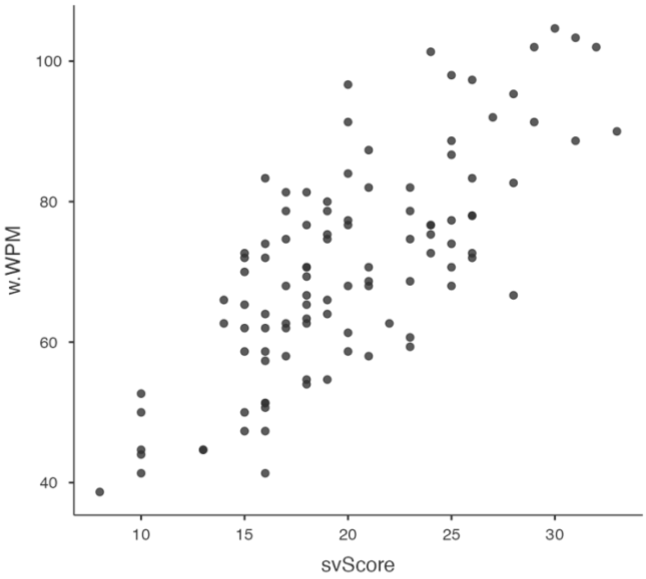


Figure 37: Scatterplot of Sentence Verification & TOWRE – pseudowords

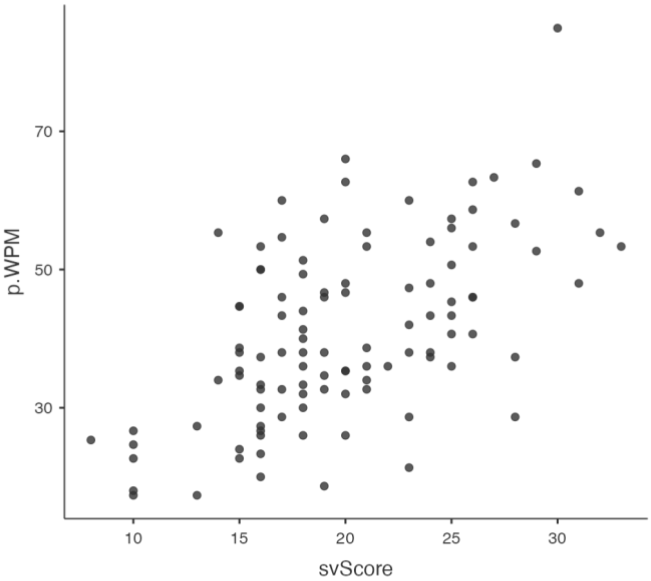


Figure 38: Scatterplot of Sentence Verification & ORF

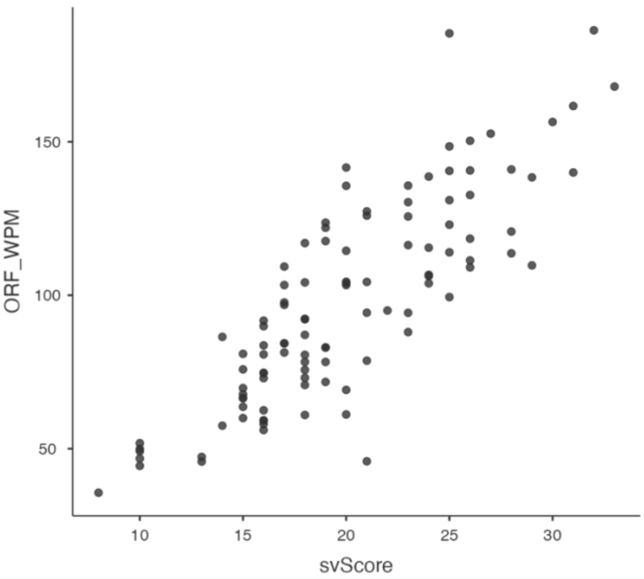


Figure 39: Scatterplot of Sentence Verification & Discrete Words-accuracy

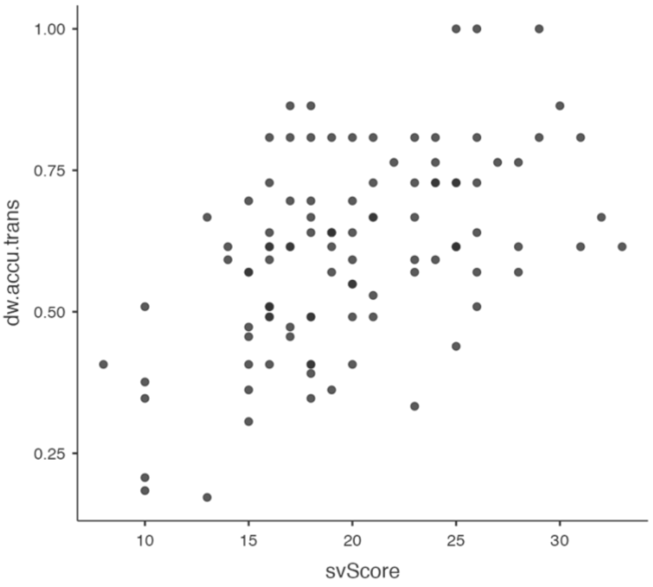


Figure 40: Scatterplot of Sentence Verification & Discrete Words-rate

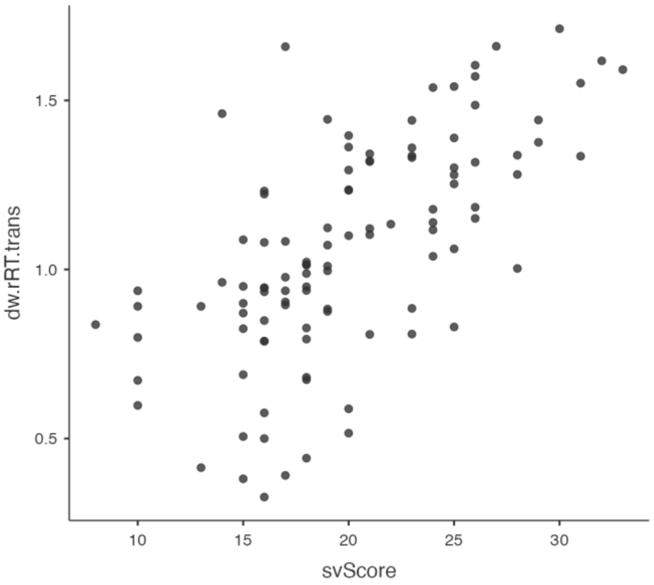


Figure 41: Scatterplot of Picture Selection & NARA

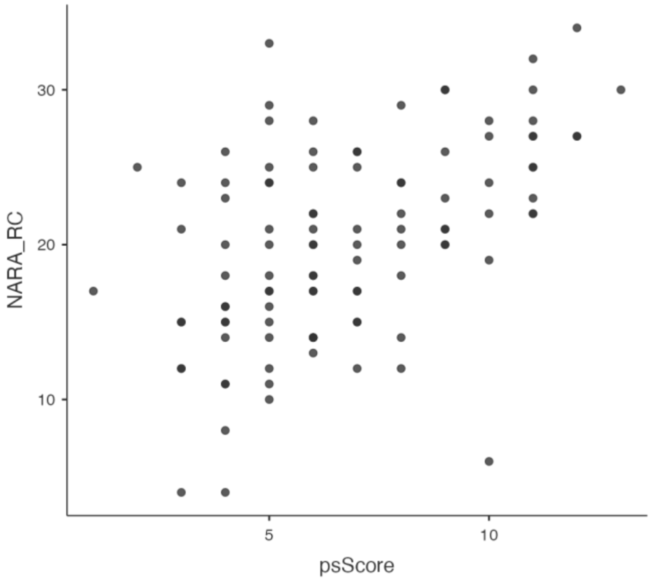


Figure 42: Scatterplot of Picture Selection & TOWRE - words

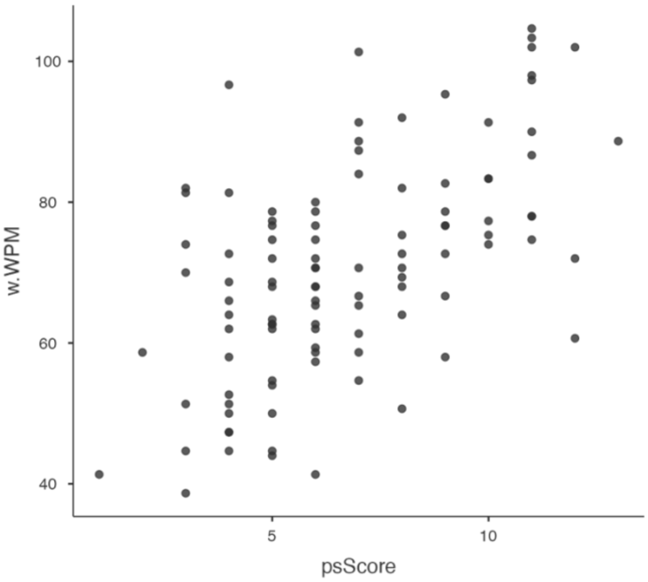


Figure 43: Scatterplot of Picture Selection & TOWRE – pseudowords

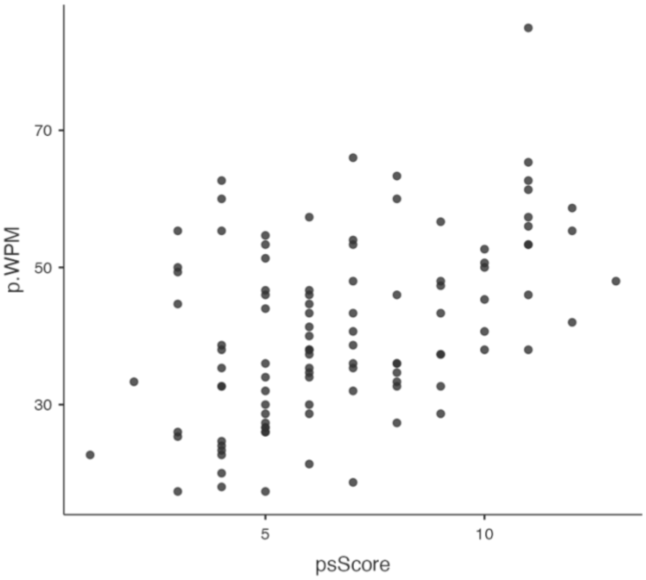


Figure 44: Scatterplot of Picture Selection & ORF

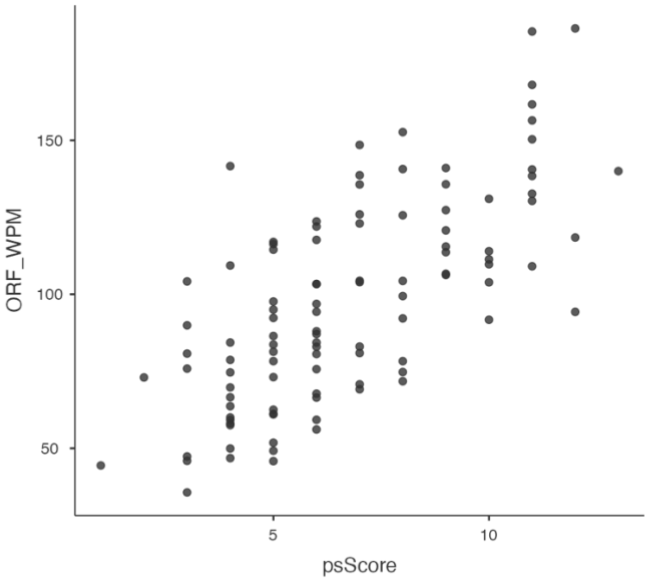


Figure 45: Scatterplot of Picture Selection & Discrete Words-accuracy

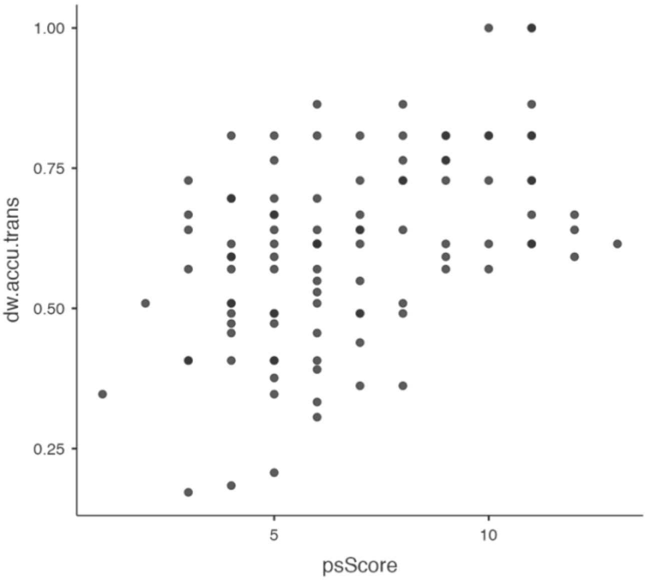
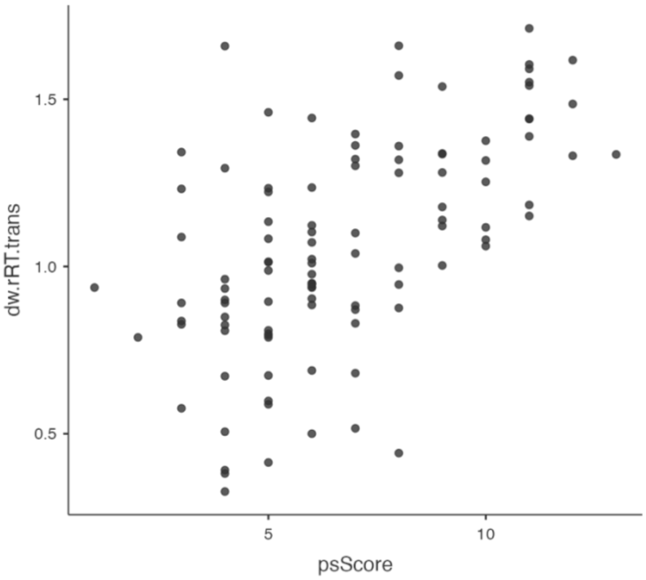


Figure 46: Scatterplot of Picture Selection & Discrete Words-rate



2.2. SCATTERPLOTS - INTERVENTION GROUP

Figure 47: Scatterplot of Sentence Verification (timepoint 1) & RAVEN

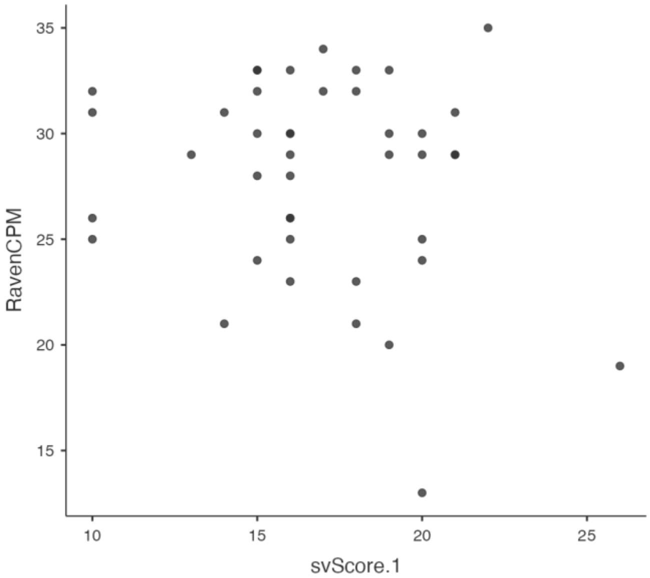


Figure 48: Scatterplot of Sentence Verification (timepoint 1) & TROG

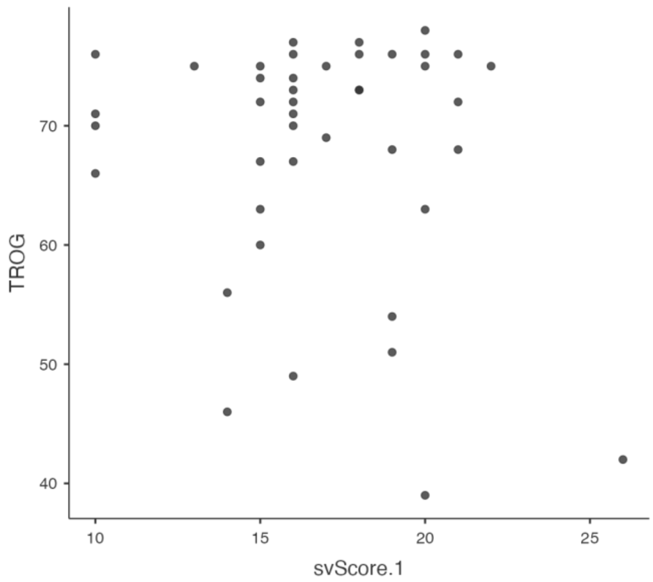


Figure 49: Scatterplot of Sentence Verification (timepoint 1) & BPVS

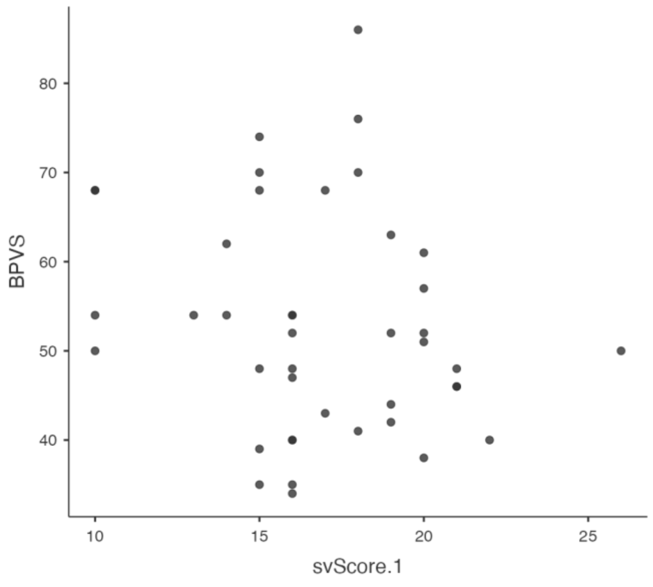


Figure 50: Scatterplot of Picture Selection (Timepoint 1) & RAVEN

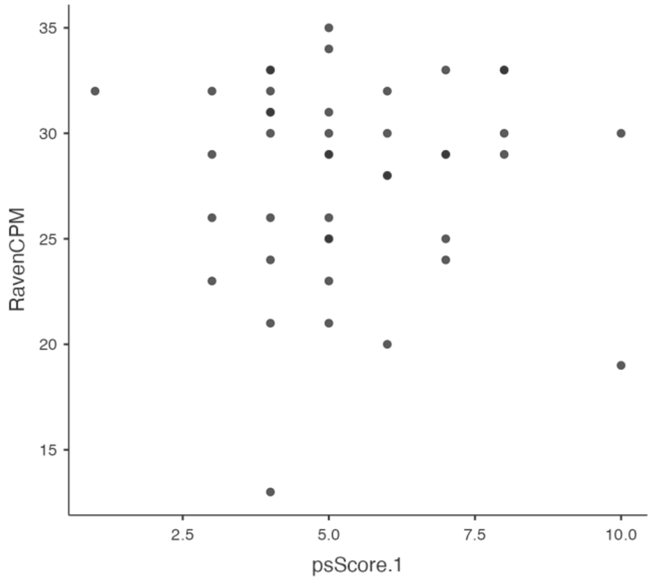


Figure 51: Scatterplot of Picture Selection (timepoint 1) & TROG

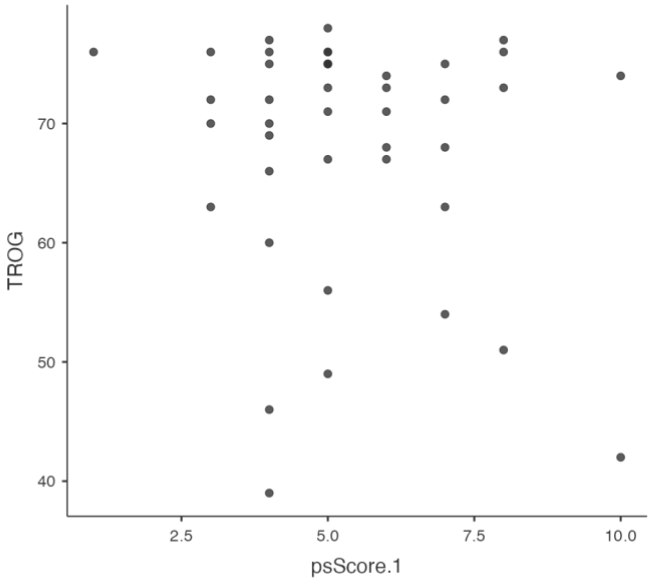


Figure 52: Scatterplot of Picture Selection (timepoint 1) & BPVS

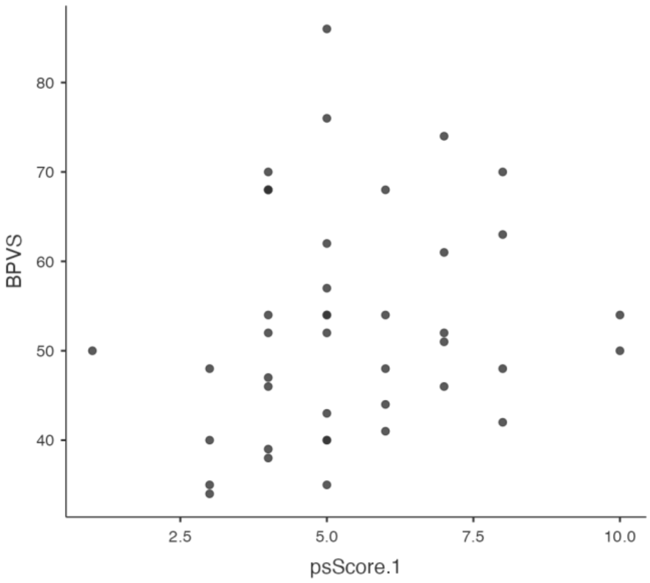


Figure 53: Scatterplot of Sentence Verification (timepoint 1) & Sentence Verification (timepoint 2)

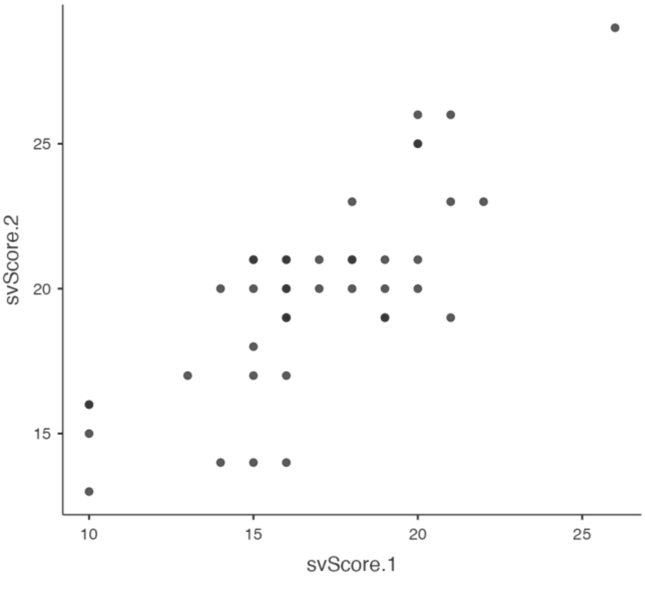
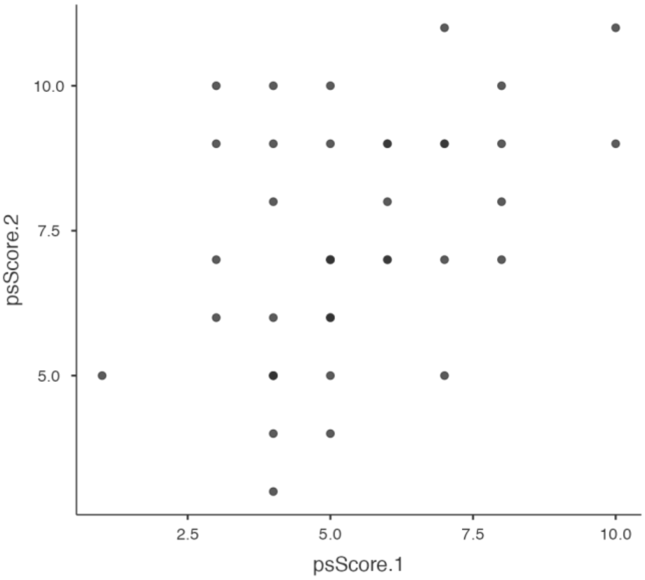


Figure 54: Scatterplot of Picture Selection (timepoint 1) & Picture Selection (timepoint 2)



Appendix 3: Information to parents and consent form

Informasjon til foreldre
om barns deltagelse i forskningsprosjektet
«BetterReading: Understanding gains in reading fluency»?

Dette er et spørsmål til foreldre om tillatelse til barns deltagelse i et forskningsprosjekt hvor formålet er *å styrke og forstå barns leseflyt*. I dette skrivet gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg og ditt barn. Dersom du er interessert i at barnet ditt skal delta ønsker vi at du fyller ut og signerer siste side, og sender denne siden tilbake med barnet på skolen.

Formål

I dag vet vi at å kunne lese flytende er avgjørende for å oppnå gode leseferdigheter. Dette betyr at for å kunne forstå og lære av tekster en leser, må lesingen være effektiv og nøyaktig med god innlevelse. Dessverre oppnår ikke alle barn optimal leseflyt, noe som fører til at de mister lesingen som et redskap til å tilegne seg ny kunnskap. Som en konsekvens vil disse barna stå i fare for å falle betraktelig bak klassekameratene sine i alle fag gjennom skoleløpet. Dette vil følgelig ha avgjørende implikasjoner for deres deltakelse i yrkeslivet og det sosiale fellesskapet forøvrig.

BetterReading er et forskningsprosjekt som skal utvikle og formidle ny kunnskap om barns leseflyt. Målet med prosjektet er å utvikle gode intervensjons- og kartleggingsverktøy som kan benyttes i skolen for å avdekke og styrke elevers leseferdigheter. Dette materialet kan komme alle elever til gode, både de som strever med lesing, og de som allerede har et godt utgangspunkt for effektiv lesing.

Hvem er ansvarlig for forskningsprosjektet?
Universitetet i Oslo, Institutt ved Spesialpedagogikk, er ansvarlig for prosjektet, ved prosjektleder og professor Athanasios Protopapas.

Det er etablert et forskningssamarbeid om prosjektet sammen med skolen barnet ditt går på.

Prosjektet er finansiert av FINNUT-programmet innen Norsk Forskningsråd.

Hvorfor får du spørsmål om å delta?
Vi ønsker å komme i kontakt med dere som er foresatte for elever på 5. trinn. Trinnet er valgt som målgruppe for prosjektet ettersom det representerer en viktig fase i barns leseutvikling, nemlig rett etter grunnleggende leseflyt er oppnådd. Som deltakere i prosjektet bidrar elevene til viktige og nyskapende funn som vil bli videreformidlet til pedagoger, foreldre og forskere gjennom blant annet foredrag og workshops. Slik kan vi øke bevisstheten om leseflyt og

hvorfor denne leseferdigheten er så viktig, og dermed jobbe mot at alle elever mottar hjelp og støtte på best mulig måte.

Hva innebærer det for barnet ditt å delta?

Hvis du velger å gi tillatelse til at ditt barn kan delta i prosjektet, innebærer det at barnet ditt blir med på en kartleggingsøkt på ca. 45 minutter. Økten gjennomføres av masterstudenter i spesialpedagogikk på et stille og skjermet rom på skolen. I løpet av kartleggingsøkten vil barnet bli bedt om å løse ulike oppgaver knyttet til lesing, som for eksempel å lese korte tekster tilpasset sin aldersgruppe, lese ordlister og frittstående ord som vises på en skjerm, og klikke på bilder som matcher ord. En del muntlige svar blir registrert automatisk ved lydopptak for videre bearbeiding og måling (f.eks. for å måle hvor lang tid det tar å gjenkjenne og uttale hvert ord).

Masterstudentene som gjennomfører kartleggingen er godt vant til å arbeide med barn som befinner seg på ulike nivå, og vil skape en trygg atmosfære hvor barnet får oppmuntring og ros hele veien. De vil starte med de enkleste oppgavene først og hele tiden vurdere vanskegraden slik at barnet opplever mestring.

Dersom dere ønsker, kan dere se på kartleggingsmaterialet på forhånd ved å ta kontakt med oss.

Det er frivillig å delta

Det er frivillig å delta i prosjektet. Hvis du velger å gi samtykke til at ditt barn kan delta, kan du når som helst trekke samtykket tilbake uten å oppgi noen grunn. Alle personopplysningene som angår ditt barn vil da bli slettet. Det vil ikke ha noen negative konsekvenser for deg eller barnet ditt hvis du ikke vil delta eller senere velger å trekke deg, og det vil ikke påvirke barnets forhold til skolen/lærer.

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger

Vi vil bare bruke opplysningene om deg og barnet ditt til formålene vi har fortalt om i dette skrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket.

Ditt og barnets navn og kontaktopplysninger vil oppbevares adskilt fra dataene (herunder kartleggingsresultater) slik at det ikke er mulig å kobles mellom person og data.

Data vil samles inn av to masterstudenter i spesialpedagogikk. Alle forskerne på prosjektet vil få tilgang til dataene. Identifiserbare data vil bli lagret på krypterte media på UiO-eide pc-er samt trygge områder innenfor UiO sitt nettverk. Lydopptak vil bli bearbeidet av UiO-forskere, og skal slettes med en gang jobben er ferdig, og senest ved BetterReading-prosjektets slutt.

Deltakerne vil ikke kunne gjenkjennes i publikasjoner eller andre kilder utarbeidet av prosjektet.

Hva skjer med opplysningene dine når vi avslutter forskningsprosjektet?

Opplysningene anonymiseres senest når prosjektet avsluttes, noe som etter planen er desember 2025.

Med en gang prosjektet er avsluttet skal alle identifiserbare opplysninger (dvs. lydopptak) slettes. Deretter skal vi offentliggjøre ikke-identifiserbare data til bruk for andre forskere ifølge «open science»-prinsipper. Anonymiserte data skal da lagres på ubestemt tid.

Dine rettigheter

Så lenge barnet ditt kan identifiseres i datamaterialet, har du rett til:

- innsyn i hvilke personopplysninger som er registrert om deg, og å få utlevert en kopi av opplysningene,
- å få rettet personopplysninger om barnet,
- å få slettet personopplysninger om barnet, og
- å sende klage til Datatilsynet om behandlingen av dine/barnets personopplysninger.

Hva gir oss rett til å behandle personopplysninger om deg?

Vi behandler opplysninger om deg og barnet ditt basert på ditt samtykke.

På oppdrag fra Institutt for spesialpedagogikk ved Universitetet i Oslo har NSD – Norsk senter for forskningsdata AS vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket (prosjektnummer 226196).

Hvor kan jeg finne ut mer?

Hvis du har spørsmål til studien, eller ønsker å benytte deg av dine rettigheter, ta kontakt med:

- Institutt for spesialpedagogikk ved Athanasios Protopapas, på e-post: athanasios.protopapas@isp.uio.no
- Vårt personvernombud, Roger Markgraf-Bye, på e-post: personvernombud@uio.no

Hvis du har spørsmål knyttet til NSD sin vurdering av prosjektet, kan du ta kontakt med:

- NSD – Norsk senter for forskningsdata AS på e-post (personverntjenester@nsd.no) eller på telefon: 55 58 21 17.

Med vennlig hilsen

Athanasios Protopapas
Professor, Institutt for spesialpedagogikk
BetterReading prosjektleder

15. september 2022

(Husk å fylle ut, signere og sende tilbake siden som følger for å delta!)

Samtykke fra foreldre om barns deltagelse i forskningsprosjektet «BetterReading: Understanding gains in reading fluency»?

Jeg har mottatt og forstått informasjon om prosjektet *BetterReading*, og har fått anledning til å stille spørsmål. Jeg samtykker til: (les og huk av alle boksene)

- □ at barnet mitt deltar individuelt i en økt på ca. 45 min der ferdigheter tilknyttet lesing blir kartlagt av en masterstudent i spesialpedagogikk
- □ at barnets muntlige svar (herunder høytlesing) blir registrert ved lydopptak
- □ at læreren kan oppgi barnets fødselsmåned og år til prosjektet
- □ at ikke-identifiserbare opplysninger lagres etter prosjektslutt på ubestemt tid
- □ at ikke-identifiserbare opplysninger blir offentliggjort anonymt etter at prosjektet er avsluttet

Jeg samtykker til at mine og barnets opplysninger behandles frem til prosjektet er avsluttet

....
(Signert av foresatte, dato)

Barnets navn:..

Kjønn: □ Jente □ Gutt

Trinn: ....

Fødselsmåned/år:...