Designing a Playful, Tablet and Group-Based Literacy Screening for German-speaking Pre-Readers: A Machine Learning Approach

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Benedikt Beuttler
aus Filderstadt

Tübingen
2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 27.11.2024

Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Jürgen Heller
2. Berichterstatterin:	Dr. Alexandra Kirsch

Acknowledgments

I want to express my sincere gratitude to my supervisors Prof. Jürgen Heller and Dr. Alexandra Kirsch, for their continuous support and guidance throughout my Ph.D. They both helped me navigate my way through new territories, gave me advice from practical experiences, and always had an open door. Thank you for allowing me to pursue my scientific interests.

Special thanks to JProf. Dr. Heiko Holz, whom I have had the privilege of calling a good friend and great advisor since my Master's studies. We worked together on many projects and I am sure many more will come. Heiko's support and help during my time as a doctoral student was immeasurably essential and valuable to me.

I would also like to take this opportunity to thank my fellow Ph.D. candidate Alisa Volkert. With her and our (almost) weekly meetings on our dissertations, we pushed each other and helped each other move forward. It was an essential contribution for me so that I would not lose focus.

A big thanks to the Research Methods and Mathematical team at the Department of Psychology, namely my former colleagues Florian Wickelmaier, Martin Losert, and Katharina Naumann. Their advice was invaluable for me in answering my questions in terms of psychological understanding and statistical analysis.

I would also like to thank Denise Löfflad, Elizabeth Bear, Niclas Trube, Heiko Holz, and Tabea Beuttler for carefully proofreading my thesis. Thank you so much for taking the time and giving me advice.

Thanks also to my parents and my two siblings. I know it has taken a while to finish this endeavor, but I am so grateful to you - dear parents - for your emotional support and for giving me confidence. You are my role models in life and I owe everything I have achieved so far to you. Thank you also to Tabea for supporting me in this journey and Samuel for paving the road to the second dissertation in the family.

Above all, I want to thank Miriam. Without her support and the necessary balance in my life, I would not have had so much perseverance in this endeavor. I am so lucky to call you my wife, and I love you!

Abstract

The ability to read and write remains highly relevant even in an age in which large language models, for example, are used to read, create, and analyze texts. The significance of properly learning how to read and write is underlined by the diagnosis of developmental dyslexia in 4 to 10% of German children, as well as by the latest IQB educational trends from 2022, which show that 30.4% of children do not meet the minimum spelling requirements. Severe problems in reading and spelling proficiencies can lead to academic, social, or personal challenges. Therefore, early diagnosis is pivotal, optimally, at the time of school enrollment. Screening for such reading and spelling shortcomings in a pre-reader age presents significant challenges due to not yet developed skills in these areas. However, several studies have highlighted diversity in the phonological perceptual domain and possible links to literacy and spelling ability even before children can read and write. Standardized pre-reader screenings that use tasks independent of written language are almost exclusively available as paper-and-pencil tests and are intended to be administered in individual sessions, thus requiring a lot of resources.

The group-based digital tool we present in this study aims to generate individualized predictors through interactive tasks, enabling the anticipation of reading and spelling abilities. We are adapting the approach of a literacy diagnostic tool for pre-readers into a digital and engaging tablet-based screening system, which, within the German-speaking domain, is mostly unexplored. First, we design the tablet and group-based application and its five screening tasks before evaluating it in a controlled field trial to extract insightful predictors. Second, we utilize machine learning techniques to analyze the study outcomes and construct prediction models. We present methods for optimizing the models and develop custom algorithms to overcome challenges present in the data. We further focus on enhancing the interpretability of regression models to provide more compelling insights into prediction outcomes. The intended result of this endeavor is to offer a solution that leverages the benefits of digitalization and addresses the unique challenges of diagnosing early-stage reading and spelling issues.

In this thesis, we investigate (1) the screening's feasibility in group settings as well as children's user and game experience with 34 German second and third graders, conduct (2) a field study with 414 German first graders with two data collection points over 1.5 years and derive (3) the predictive power of five digitized tasks for children

of pre-reader age using (4) machine learning models from that study. The tablet data from the first data collection point of the field study is used to predict reading and spelling performance at the second data collection point. The results for (1) showed that feasibility for group sessions with up to 10 children was achieved through optimized seat arrangements, specific materials, and engaging (task-)elements to motivate children and minimize distractions. As a result of the study (2), we found several tasks whose features are significant predictors of literacy skills, while others failed to differentiate groups adequately (3). Our developed algorithm for (4) assessed various models and oversampling rates and eventually identified Random Forest as the superior model for predicting reading and spelling skills. Once the results are classified and compared to other traditional literacy screenings, our screening approach performs moderately well with a balanced accuracy of 72.5% and a RATZ index of 46%. Incorporating prediction intervals has enhanced result precision and interpretability.

Zusammenfassung

Die Fähigkeit, lesen und schreiben zu können, ist auch in einem Zeitalter, in dem Sprachmodelle eingesetzt werden, um Texte zu lesen, zu erstellen und zu analysieren, von großer Bedeutung. Wie wichtig es ist, richtig lesen und schreiben zu lernen, zeigen die Diagnose einer Entwicklungsdyslexie bei 4 bis 10% der Kinder in Deutschland und die aktuellen IQB-Bildungstrends 2022, die zeigen, dass 30,4% der Kinder die Mindestanforderungen in der Rechtschreibung nicht erfüllen. Schwerwiegende Lese- und Rechtschreibprobleme können zu schulischen, sozialen und persönlichen Problemen führen. Eine frühzeitige Diagnose ist daher von großer Bedeutung, am besten vor oder direkt nach der Einschulung der Kinder. Das Screening solcher Lese- und Rechtschreibschwächen im Vorlesealter stellt aufgrund der unterentwickelten Fähigkeiten in diesen Bereichen eine große Herausforderung dar. Mehrere Studien haben jedoch die Vielfalt im phonologischen Wahrnehmungsbereich und mögliche Zusammenhänge dieser mit der Lese- und Rechtschreibfähigkeit, ohne speziell lesen und schreiben zu können, aufgezeigt. Standardisierte Screenings für Leseanfänger, die schriftsprachunabhängige Aufgaben verwenden, sind fast ausschließlich als analoge Tests verfügbar und sollen in Einzelsitzungen durchgeführt werden, was für das Lehrpersonal einen hohen Ressourcenaufwand bedeutet.

Wir adaptieren den Ansatz eines Diagnostik-Tools für Leseanfänger in ein digitales und ansprechendes tablet-basiertes Screening-System, das im deutschsprachigen Raum noch weitgehend unerforscht ist. Dieses gruppenbasierte digitale Tool zielt darauf ab, durch interaktive Aufgaben individualisierte Prädiktoren zu generieren, die das Vorhersagen von Lese- und Rechtschreibfähigkeiten ermöglichen. Zunächst konzipieren und evaluieren wir die tablet- und gruppenbasierte Anwendung mit fünf Screening-Aufgaben anhand einer entsprechenden Feldstudie, die die Gewinnung aufschlussreicher Prädiktoren ermöglicht. Des Weiteren setzen wir Techniken des maschinellen Lernens ein, um die Ergebnisse der Studie zu analysieren und Vorhersagemodelle daraus abzuleiten. Wir stellen Methoden zur Optimierung der Modelle vor und entwickeln benutzerdefinierte Algorithmen, um die in den Daten vorhandenen Herausforderungen zu bewältigen. Zu diesem Zweck entwickeln wir optimierte Vorhersagemodelle und benutzerdefinierte Algorithmen, um die in den Daten vorhandenen Herausforderungen zu bewältigen. Darüber hinaus konzentrieren wir uns auf die Verbesserung der Interpretierbarkeit der Modelle, um überzeugendere Einblicke in die Vorhersageergebnisse zu ermöglichen. Das

angestrebte Ergebnis dieser Arbeit ist eine innovative Anwendung, die die Vorteile der Digitalisierung nutzt und die besonderen Herausforderungen der Früherkennung von Lese- und Rechtschreibproblemen angeht.

In dieser Arbeit untersuchen wir (1) die Durchführbarkeit des Screenings im Gruppensetting sowie die Nutzungs- und Spielerfahrungen der Kinder mit 34 deutschen Zweit- und Drittklässlern, führen (2) eine Feldstudie mit 414 deutschen Erstklässlern mit zwei Erhebungszeitpunkten über 1,5 Jahre durch und leiten (3) die Vorhersagekraft von fünf digitalisierten Aufgaben für Kinder im Vorlesealter mit (4) maschinellen Lernmodellen aus dieser Studie ab. Die Tablet-Daten aus dem ersten Erhebungszeitpunkt der Feldstudie werden zur Vorhersage der Lese- und Rechtschreibleistung zum zweiten Erhebungszeitpunkt verwendet. Die Ergebnisse für (1) zeigten, dass die Durchführbarkeit für Gruppensitzungen mit bis zu 10 Kindern durch optimierte Sitzanordnungen, spezifische Materialien und ansprechende (Spiel-)Elemente erreicht wurde, um die Kinder zu motivieren und mögliche Ablenkungen zu minimieren. Als Ergebnis der Studie (2) fanden wir mehrere Aufgaben, deren Merkmale signifikante Prädiktoren für Lese- und Schreibfähigkeiten sind, während andere die Gruppen nicht ausreichend differenzieren konnten (3). Der von uns entwickelte Algorithmus für (4) analysierte verschiedene Modelle und Oversampling-Raten und identifizierte schließlich Random Forest als das beste Modell für die Vorhersage von Lese- und Rechtschreibfähigkeiten. Nach der Klassifizierung der Ergebnisse und dem Vergleich mit anderen traditionellen Lese- und Schreibfähigkeits-Screenings schneidet unser Ansatz mit einer Genauigkeit von 72,5% und einem RATZ-Index von 46% ähnlich gut ab. Die Einbeziehung von Vorhersageintervallen hat die Präzision und Interpretierbarkeit der Ergebnisse verbessert.

Table of Contents

Part A

Background

Chapter 1

Introduction and Motivation

Despite the increasing digitization in classrooms and the use of the latest AI models, such as OpenAI's GPT-4o and Anthropic's Claude to generate text, reading and writing are still essential skills that children and young learners have to acquire. Unfortunately, approximately 30.4% of German children fail to meet the minimum spelling standard according to the recent educational trend (Stanat et al., 2023) by the German Institute for Educational Quality Improvement (Institut zur Qualitätsentwicklung im Bildungswesen), and between 4% and 10% of German children struggle to achieve proficiency in reading and spelling, leading to a diagnosis of developmental dyslexia or severe literacy difficulties (Moll et al., 2014; Moll and Landerl, 2009). Dyslexia is one of the most common psychological disorders among children and adolescents in Germany (Schulte-Körne, 2010). Difficulties with reading and spelling have a negative impact on many elements of a child's life, including personal development (Schulte-Körne, 2010), social interactions (Beddington et al., 2008), and academic success (Daniel et al., 2006). Dyslexic children face considerable challenges in learning literacy skills, which can reduce their motivation to learn and decrease their confidence in achieving language competence (Bender et al., 2017). Furthermore, children are more vulnerable to negative thoughts, despair and anxiety regarding their school life (Schulte-Körne, 2010). As a result, interventions and therapies for dyslexic children, or children with weak literacy skills in general, are critical and should begin as soon as possible. Research has shown that early intervention has a lasting impact on learning success (Marx and Lenhard, 2011; Rauschenberger et al., 2018b). Therefore, early diagnosis of reading and spelling weaknesses is the first and most crucial step in supporting affected children. However, precisely diagnos-

ing children's literacy inadequacies, especially at a pre-reader age, poses a substantial challenge. Traditional reading, writing, or phonemic awareness assessment, which are well-established metrics for predicting reading and spelling progress (Moll et al., 2012), are not viable options at this point because these skills are still not appropriately developed (Hesketh et al., 2007). As a result, language-independent metrics such as naming speed (Landerl et al., 2019) or auditory perception (Maurer et al., 2003) are used in pre-reader screenings. There are paper-and-pencil screenings in the area of German spoken language that attempt to classify pre-reader children into dyslexic and non-dyslexic. Still, the prognostic validity is often insufficient (see Marx and Lenhard, 2011, for an overview). Digitizing such an approach and using the increasing use and availability of digital devices in schools (Forsa, 2024) to create an engaging, motivating, and group-based system with all the benefits of digitization (e.g., automatic scoring, standardizing the procedure, accurately measuring reaction times) has been explored little in research focusing on the German language.

1.1 Aims and contributions

This thesis aims to design and implement a tablet-based literacy screening for pre-readers that is digital, playful and group-oriented. Figure 1.1 shows the process to achieve that goal and how we build predictive models for that purpose. In such a screening, individual predictors, e.g., performance indicators, are computed for each child from the (performance) data of evidence-based tasks, allowing the prediction of their reading and spelling abilities.

We also aim to assess whether the screening can be successfully carried out independently by children in a group and whether the game design delivers a playful experience. We split this thesis into two main parts, with two overarching goals in mind: (1) designing and creating our tablet-based screening application as well as conducting a field study with it to ultimately derive meaningful predictors from the screening tasks and (2) analyzing the resulting predictors using machine learning methods to generate optimized prediction models to predict literacy skills.

Following this, we formulate two research objectives: designing and developing a group-based literacy screening for pre-reader children and incorporating machine learning for predicting literacy skills.

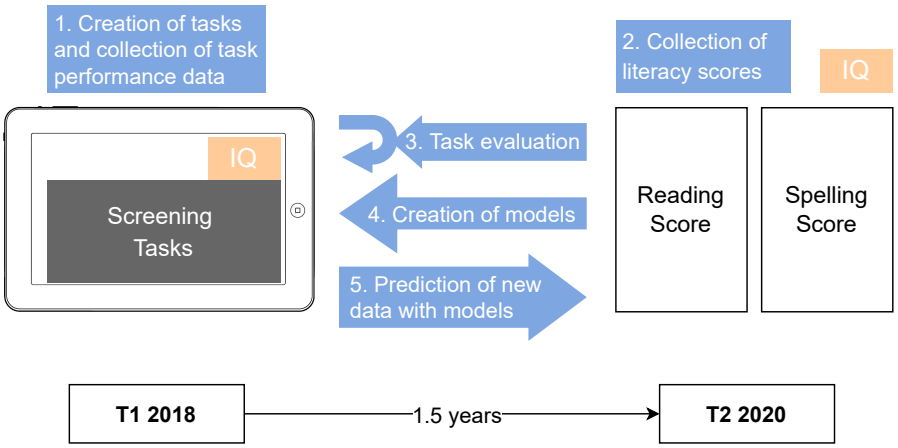


Figure 1.1: Workflow of this thesis from task generation and data collection to the creation of prediction models. We first (1) create screening tasks and collect the performance data of children that just have started school, (2) collect the literacy data 1.5 years later when the children are in second grade, (3) evaluate the screening tasks and (4) use this information to create machine learning models for literacy score prediction. These models could then be used for (5) screening evaluation and future prediction of children's literacy scores.

1. Design and development of a feasible, group-based literacy screening and derivation of relevant predictors.

   RQ1.1. *Can children use the screening independently and how do they perceive it?*

   For the screening to be feasible in a group setting, it has to be simple to understand and use by children of pre-reader age. It also needs to offer a good game and user experience to engage and motivate children to complete the screening.

   RQ1.2. *Is the presented literacy screening applicable in a group setting?*

   In general, group testing can help teachers to save valuable time. In order to carry out the screening in a group environment, an application that is easy for children to understand and that enables focused work in such an environment is needed.

   RQ1.3. *How can the screening tasks be implemented effectively in a digital, game-based environment?*

Some screening tasks are already used with pencil and paper in therapies or schools. We investigate the implementation of these tasks and all other tasks for tablet devices, and we look at the unique features that emerge or are required by this change in medium.

RQ1.4. *Are the screening tasks theoretically sound and which predictors best predict literacy skills?*
The meaningfulness of the tasks is the most important building block in the screening. Which metrics can be tracked and how do they relate to reading and spelling skills? The tasks therefore form the basis for all further analyses and model training.

2. Incorporating machine learning methods and algorithms to improve model predictions of literacy skills.

RQ2.1. *How can the biased data from the screening be processed so that it is suitable for prediction?*
Regarding task performance, especially in reading and spelling, the data is usually biased, e.g. unevenly distributed. Depending on whether the problem is defined as a classification or a regression problem, there are different approaches to address these data biases.

RQ2.2. *Which learning algorithm best predicts reading and spelling skills based on our screening data?*
Given the particular data structure, some algorithms can handle such biases better or worse. In combination with machine learning methods, such as cross-validation, hyperparameter tuning or resampling, the question arises which model predicts most effectively.

RQ2.3. *How well does our screening approach compare to conventional ones?*
To enable a comparison between our digital, group-based screening and previous German dyslexia screenings, model results need to be classified and specific metrics calculated. Can the screenings be compared at all?

RQ2.4. *What is the most meaningful way to interpret the results of the prediction models?*
A common interpretation of the results of the common dyslexia screenings only allows a division into two classes - dyslexic and non-dyslexic.

Instead, we try to formulate our approach as a regression question with meaningful interpretations of the results.

The above research questions call for a unique approach to literacy screening incorporating machine learning We make the following overarching contributions to research on digital literacy screenings:

- Creating predictive tasks for children of pre-reader age. On the one hand, existing analog tasks are translated appropriately for the digital domain, and on the other hand, new tasks are created exploratory based on the latest research findings.
- Integration of motivational and gamified elements for a feasible group screening, which promotes independent work on the tasks. Paired with an innovative in-game questionnaire, we also evaluate the feasibility of the screening.
- Application of machine learning methods for training and evaluating literacy screening models, given the specific dataset at hand. We translate the usual approach to literacy screening from a classification problem to a regression problem and present a custom algorithm for data and model optimization.

1.2 Reader's guide

This readers guide provides an overview of how this thesis is structured and what you can expect to find in each part and chapter.

Part A We first introduce the reader to dyslexia, its causes, its connection to weak literacy skills according to our definition of it and the related work of dyslexia screenings in all of their forms - traditional, digital and game-based (Chapter 2). We provide an overview and comparison of traditional German dyslexia screenings and look at modern approaches to dyslexia screenings relevant to us. Finally, in Chapter 3, we look at what makes serious games, that can be implemented in game-based learning and assessment, so successful and how their integration into an assessment approach enables independence within group-based testing.

Part B In Chapter 4, we discuss individual testing in a group setting, focusing on our approach to adapt the screening for group testing while maintaining the integrity of the individual assessment process. Additionally, we highlight the game elements incorpo-

rated to ensure a successful and engaging execution of the screening. Next, we present an overview of all the tasks included in the screening process and their digital implementation. We delve into their theoretical background, procedure and construction, providing a clear understanding of their relevance to literacy screening.

As we will argue, group testing is, in many ways, less resource-intensive and more efficient in its application, but its feasibility is yet to be explored. We therefore evaluate and discuss the game and user experience based on a pilot study in Chapter 5. For this purpose, we briefly describe the study methods with the integration of the questionnaires before evaluating the feasibility of the screening process.

In Chapter 6, we describe the design, execution, and analysis of a large-scale study with 414 first-grade children. We explain the participant recruitment, methodology, measurements, data collection procedures and statistical analysis methods employed to investigate the screening tasks. Finally, we present and discuss the results obtained from the main study in Chapter 7, highlighting significant observations, comparing them to previous findings and discussing possible shortcomings in the tasks. Most importantly, we select predictors by statistically analyzing each task and identifying the parameters best suited for predicting reading and spelling skills.

Part C This part discusses machine learning and its practical applications in our screening. We explore the challenges of imbalanced datasets and various strategies to address this issue. In addition, we provide insights into the requirements for training and evaluating a machine learning model and discuss the evaluation of our screening as a whole to make it comparable to other, traditional screenings. To enhance the informative value of our regression models, we use prediction intervals that consider the inherent uncertainty present in the prediction results.

In the practical part (Chapter 10), we compare and evaluate the performance of different models, enabling us to identify the best-performing model and uncover the reasons behind its success. We further discuss the prediction intervals and look at a comparison to traditional screening.

Part D Finally, we summarize the results of the studies conducted and the data analysis, provide recommendations for the implementation of group-based literacy screenings and discuss our research questions. Lastly, we also take a look at the future of digital literacy screenings.

Chapter 2

Identifying Dyslexia

2.1 Causes of dyslexia

One of the most important skills to be acquired in a child's early development is the ability to read and write properly. If possible shortcomings in these skills are not adequately diagnosed and treated in time, they can impair the academic (Daniel et al., 2006), personal (Ise and Schulte-Körne, 2010; Mayer, 2016) and social (Beddington et al., 2013) development of the child in the short and long term.

In recent decades, science has shown that the development of a nation and the use of its mental capital is as important of a factor for its economic competitiveness and prosperity as it is for the mental health, well-being and social integration and inclusion of its citizens (Beddington et al., 2013). The ability to read and write is a fundamental capability for an individual growing up. Failure to acquire a written language and learning deficits can have a negative impact. A resulting lack of motivation to continue learning in general and the loss of self-confidence to develop a comprehension of literacy language can hinder a child's development (Bender et al., 2017). Learning disorders can also negatively impact mental health, social and cultural participation, and educational and personal development (Beddington et al., 2013; Bender et al., 2017; Daniel et al., 2006). A higher chance of dropping out of school, resulting in a bad employment outlook or unemployment, can be the consequences (Daniel et al., 2006; Esser et al., 2002). For this reason, the timely diagnosis and intervention of weaknesses in the acquisition of the written language is a high priority in social and educational policy.

2.1.1 Definition dyslexia and the distinction to literacy weakness

According to the World Health Organization's international classification of diseases (ICD-10, Chapter V (Dilling et al., 2015)), a reading and spelling disorder is classified as a circumscribed developmental disorder of school skills. It is characterized by an impairment of reading and spelling skills, which cannot be attributed, among other things, to a lack of learning opportunities, such as inadequate schooling or instruction, a general reduction in intelligence, deficits in hearing and vision, or a neurological, psychiatric or other disorder (Mayer, 2016; Schulte-Körne, 2015). A reduced reading speed primarily marks weaknesses in reading behavior, numerous reading errors (e.g., omission, transposition or replacement of parts of words or entire words) and reduced reading comprehension. A spelling weakness, on the other hand, is characterized by a high number of spelling errors (e.g., mixing up letters in the wrong order, omitting phonemes or words) and error inconsistency. There are conceptual distinctions in the type of dyslexia. The disease *dyslexia*, according to the WHO, is an isolated performance disorder with specific difficulties in acquiring written language. Dyslexic weaknesses, however, are unspecific difficulties in acquiring written language. Because this distinction is not essential for the diagnostic prognosis and to ensure a better reading experience, the term dyslexia will be further used for both cases, with less focus on dyslexia as a specific disorder, or completely be replaced by the term *literacy weakness*. For the definition of who is considered a weak reader or weak speller, we follow the approach of other research to identify low-performing children in an educational context, such as Mayringer and Wimmer (2014). In it, the authors only include data from children who reported German as their native language and who scored less than one standard deviation below the mean on the raw scores for the respective literacy skills.

The causes of dyslexia have not yet been fully clarified. The most common assumption is that the causes are located at multiple levels and that neurobiological factors are of particular importance, leading to difficulties in phonological information processing at the linguistic-cognitive level. These include processes related to the recognition, holding, manipulation, storage and recall of linguistic units (Lindberg, 2016). The construct assigned to phonological information processing and decisively associated with the development of dyslexia is called phonological awareness. It describes the insight into the sound structure of speech (Mayer, 2016). From an empirical perspective, environmental

factors, such as parental influences, are attributed a much smaller role than is usually assumed (Schulte-Körne and Remschmidt, 2003a).

2.1.2 Prevalence

Unfortunately, it is estimated that about 5-10% of the world's population has reading and writing difficulties (Dyslexia International, 2014), where some evidence even suggests the figure to be around 17% (Sprenger-Charolles et al., 2011). It is considered to be the reason for approximately 80% of all learning disabilities (Vidyasagar and Pammer, 2010). With approximately 4-10% of German children suffering from the aforementioned shortcomings (Moll et al., 2014; Moll and Landerl, 2009), development dyslexia is considered one of the most common child and adolescent psychiatric diseases in Germany (Schulte-Körne and Remschmidt, 2003b). Since reading is a fundamental prerequisite for acquiring knowledge in almost all school-related areas, children with dyslexia often stand out due to poor performance (Mayer, 2016). Compared to their peers with similar cognitive capabilities, they also achieve lower levels of education and are more likely to be unemployed (Esser et al., 2002).
Treatments and interventions are usually more effective the earlier they can be implemented (Marx and Lenhard, 2011; Rauschenberger et al., 2018b). Identifying developmental dyslexia as early as possible is therefore the first and most crucial step in a successful therapy.

2.2 From analog to digital, game-based dyslexia screenings

In the last 20 years, many screening procedures for early prediction of reading and writing have been developed, primarily based on research results on phonological information processing and phonological deficit theory. This well-developed and evidence-based theory sees a causal role of phonological skills in children's development of reading and spelling (Ramus, 2003). Children with good phonological skills become good readers and spellers, while children with poor phonological skills progress more poorly (Goswami, 1999). As such, deficient phonological awareness is known as one major cause of dyslexia (Bradley and Bryant, 1983; Snowling, 1995).

It's not trivial to correctly select children who require treatment at an early develop-

mental age. Since their reading and spelling errors mainly characterize dyslexic children, different indicators are required for pre-readers at the end of kindergarten or the first year of school. Children at this age have already acquired the necessary precursor skills for reading and spelling. Testing for these skills is therefore a common method in current screenings that aim to identify children at risk of dyslexia. This process - called selective prevention - requires a valid classification of children that need further treatment (Marx and Lenhard, 2011). In contrast to selective prevention, a universal prevention approach does not distinguish between groups and applies treatment to all children. Although there are benefits to the universal approach (e.g., no stigmatization, ease of use), the drawbacks outweigh (e.g., high personnel and financial costs, children in need don't get suitable, extended treatment).

Reliably predicting the development of written language at a pre-reader age is made more difficult by the fact that the type of tasks these children are able to perform is limited as well as the attention they have (Jung et al., 2021). In addition to cognitive and motor limitations, there are limitations in the early years of development, especially in auditory and phonological perception, which is undoubtedly closely related to reading and spelling development (Maurer et al., 2003). For example, awareness of syllables or rhymes is already developed at the age of 42 months (Dodd et al., 1989). Still, awareness of phonemes seems to develop only later with the acquisition of the written language (Hesketh et al., 2007). Accordingly, screenings for pre-readers test, in particular, the abilities of phonological awareness in a broader sense (e.g., syllable segmentation or rhyme recognition). However, reading and spelling development cannot be predicted sufficiently well based on these skills. More suitable predictors would be language-relevant indicators in the field of phonological awareness in the narrower sense (such as phoneme recognition or manipulation). These skills contribute considerably to the prediction of written language development but are only getting relevant when learning how to read and write starting second grade (Moll and Landerl, 2009; Moll et al., 2012). Another problem is that those predictors generally produce floor effects at pre-reader age and do not differentiate well enough in the lower performance range (Marx and Lenhard, 2011).

2.2.1 Analog ways of screening dyslexia for pre-readers

At the time of writing this dissertation, mostly analog screenings are established in the German educational system. In the following, we list these screenings and their

advantages, disadvantages and general differences, taken mostly from the meta-study by Marx and Lenhard (2011). Table 2.1 provides a summary overview.

DP: This is one of the oldest analog screenings and, years later, was part of a conversation about its validity and the methodologies used (Dummer-Smoch et al., 2012). The DP was developed before research on phonological information processing (1980-1990) and therefore does not have indicators of this type. As mentioned before, this type of processing is an important component in the development of language.

PB-LRS: Compared to other screenings, the spelling skills in the PB-LRS screening are tested early - at the end of the first class. When compiling the PB-LRS data, there may have been an unintentional bias because, for example, the teachers recorded information about the children and not the parents (e.g., information on mother tongue) and the number of children at risk after the reading tests is lower at the second measurement point than at the first. This is due to the fact that some at-risk children were already placed in support classes after the tests at the first measurement point. On the positive side, the PB-LRS is one of the few German literacy screenings that includes tests on phonological awareness in the narrow sense and allows for group testing.

RdH: *Rundgang durch Hörhausen* is one of the first screenings that has a rather playful structure and wraps the testing in a story. The screening can be conducted in both individual and group settings, although conducting it in a group setting is quite time-consuming due to the large amount of materials and preparation involved.

BISC: The BISC should be used twice: the first ten months and the second four months before enrolling in school. The fact that it is time- and therefore age-dependent significantly limits its applicability. Furthermore, the manual does not document how the risk areas were calculated or on what basis they were determined.

MÜSC: The MÜSC is designed as a group test, where only a maximum of 8 children can be tested at the same time. The test consists of two parts, whereby the second part can also be carried out on a different day but within the first five weeks after starting school.

HASE: The HASE screening's validation process assessed prediction accuracy using collected data. A follow-up study is essential for comprehensive validation metrics to

test the screening with new data and prevent overfitting.

DESK 3-6 and DESK 3-6 R: This screening aims at children of kindergarten age. Teachers carry out the developmental assessment, which is relatively broad and, in principle, not specifically designed to predict dyslexia. Tasks requiring fine motor skills, gross motor skills, cognition, social development and language are also performed.

WÜSC: According to the authors, since 2020, the WÜSC is the continuation of the BISC and recommended instead. Like the BISC, normative values are available for two periods before school enrollment, one 10-11 months in advance and the other 4-5 months in advance. Compared to the other screenings, for the development and validation of the screening, the WÜSC had a low sample size of $N = 330$ children. Due to the early timing of testing, the WÜSC, like the BISC, relies on educators for ratings and assessments during test administration.

In their meta-study, Marx and Lenhard (2011) reviewed objectivity, reliability and validity of aforementioned German analog dyslexia screenings. According to the authors, objectivity is generally given except for the DESK 3-6 test. Data for calculating reliability are only available for a few screenings. The authors note that some screenings introduce an intermediate category in the classification of children due to their uncertainty. The so-called edge-case class is intended to cover children who cannot be classified in either the dyslexic or the non-dyslexic group. Concerning prognostic validity, according to the authors and the numbers available, none of the mentioned screenings allow confident predictions.

Recently, game-based approaches have gained popularity due to their ability to reduce the high costs and extensive resources needed for individual testing while also engaging participants more effectively. This aspect, combined with the fact that some screenings still need manual assessment by educators, which can result in a lack of objectivity and standardization, leads to a more technical and digitized approach to screen for reading and spelling difficulties.

2.2.2 Digital-based screenings

When we look at the landscape of digital and game-based screening for dyslexia, especially in Germany, we see less choice compared to the analog alternatives. From a

Table 2.1: List of German, analog dyslexia screenings for children in a pre-reader age.

Screening		Authors	Year	Tasks	Age	Test type	Duration	Scale values	Data	Dyslexic
DP	Differenzierungs-probe	Breuer und Weuffen	1975	Exercises on the ability to differentiate in five different areas of perception	4-7	Individual test	7 min.	Separate evaluation of all areas of perception	Study on the progn. validity of DP Steinbrink et al. (2010) [a]; N=664	—
PB-LRS	Gruppentest zur Früherkennung von LRS	Barth and Gomm (2004b)	2004	Measurement of performance in the field of phonological awareness (in a narrow and broad sense)	5-6	Group test	60 min.	Summarized score	Barth and Gomm (2004a); N=450	15%
RdH	Rundgang durch Hörhausen	Frank et al. (2001)	2002	Measurement of performance in the field of phonological awareness (in a narrow and broad sense)	5-6	Individual test & group test	45 min.	Summarized score with correlations to teacher assessment	Einsiedler et al. (2002), N=375	20%
BISC	Bielefelder Screening	Jansen et al. (2002)	1999	Measurement of performance in working memory, (visual) attention and phonological awareness in the broader sense	5	Individual test	25 min.	Risk points per exercise	Replication shows significantly lower prognostic validity Marx and Weber (2006)	15%
MÜSC	Münsteraner Screening	Mannhaupt (2006)	2006	Exercises on phonological awareness in a broader sense, short-term memory for speech and visual attention	5-6	Group test	2x25 min.	Risk points per exercise	Mannhaupt (2006); N=2896	—
HASE	Heidelberger Auditive Screening in der Einschuluntersuchung	Schöler and Brunner (2008)	2007	Tasks to record the phonological working memory and the state of language development	4-6	Individual test	10 min.	Risk points per exercise	Prüfung der Validität durch Neugebauer and Becker-Mrotzek (2013) und Treutlein et al. (2011)	40%
DESK 3-6	Dortmunder Entwicklungsscreening für den Kindergarten	Tröster et al. (2004)	2004	General identification of developmental and behavioral problems (e.g. motor and social developments). Additional data collection by the educators	5-6	Individual test	3-4 weeks	Summarized score with correlations to teacher assessment	Tröster et al. (2004); N=1492 (3-6 jährige)	11%
DESK 3-6 R	Dortmunder Entwicklungsscreening für den Kindergarten	Tröster et al. (2016)	2016	General identification of developmental and behavioral problems (e.g., motor and social developments). Additional data collection by the educators	5-6	Individual test	3-4 weeks	Summarized score with correlations to teacher assessment	Tröster et al. (2016); N=1693 (3-6 jährige)	11%
WÜSC	Würzburger Screening	Endlich et al. (2019)	2020	Exercises on reading speed, working memory and phonological awareness in a broader sense	5	Individual test	25 min.	Weighted summarized score	Endlich et al. (2019); N=192	—

[a]The study on prognostic validity by Steinbrink et al. (2010) and the methodologies used in the study were the basis for several discussions (Dummer-Smoch et al., 2012)

practical standpoint, digitization can improve any screening procedure by providing a way to standardize and objectify the approach.

As early as 2009, there was research in the field of literacy screenings that discussed and tested automated data collection from preschool children. The authors digitized standardized tests and made voice recordings of the children (Bocklet et al., 2009). The authors focused on the theory and practice of data collection from the young target audience. However, for a possible practical application in a screening tool, the data would still have had to be processed manually.

In order to standardize the screening procedure, some German analog dyslexia screenings started including digitally implemented elements, such as the read-aloud instructions and explanations via CD in the BISC or the computer-aided version of the HASE screening.

The HASE test is the only German dyslexia screening from the aforementioned Table 2.1 that allows for a completely computer-assisted testing procedure. The underlying tasks, which mostly include an auditory component, are well-suited for digitization. Nevertheless, the evaluation is done manually. We are not aware of any German screening specifically for dyslexia that is digitized from data collection to evaluation. Systems like the AGTB 5-12 (Hasselhorn et al., 2012) and the Münsteraner Rechtschreibanalyse (MRA) (Schönweiss, 2007) offer both a computer-based test and an automated analysis. The AGTB 5-12 aims to screen pre-readers from 5-12 years old primarily regarding their working memory, which also includes testing for phonological processing. With 80-90 minutes to complete the test, it can be taxing for young children. The MRA screening, which is offered to first- and second-graders at the earliest, queries children's reading and spelling performance via a fill-in-the-blank text and tests for weaknesses in reading-spelling acquisition. Parents can conduct the MRA from home via a website and get immediate results on their children's performance. However, this system is not suitable for pre-reader children.

2.2.3 Digital game-based screenings

When we look into game-based dyslexia screenings, we see very few approaches in the German landscape. Barth and Gomm (2004b) implemented a story-driven approach in the PB-LRS where a dwarf leads the children through the test. Similarly, at the RdH, the children explore the village of Hörhausen and are provided with playful, hand-

held elements such as a homemade mailbox when conducting the tests. However, these screenings are only available in the analog version.

In the German-speaking area, there have been few approaches to implementing a dyslexia screening in digital form and with playful elements. Rauschenberger and colleagues have been researching in this direction since 2016 and have presented a concept with *MusVis* (formerly *DysMusic*), in which possible dyslexic pre-readers are to be detected at an early stage using exercises with auditory and visual elements. Language-independent item development and use were given a lot of attention. The approach still showed potential in the development of the screening, e.g., in the explainability of the machine learning models and the integration and implementation of the gamified elements (Rauschenberger et al., 2022, 2018a, 2017).

Research in the direction of digitized, game-based dyslexia screenings therefore still seems to be in its infancy, which is why it might be worth looking at other countries.

Researchers from Malaysia have piloted a pre-reader literacy screening for the Malay language, called *Dleksia*. The main goal was to test the usability of a handheld device and how the children perceived the assessment. When conducting the mobile game, approximately 82% of the subjects did not feel they were being tested ($n = 11$) (Mohtaram et al., 2017).

Another study in Italy conducted a screening consisting of three serious games that present children with cognitive tasks like problem-solving, image discrimination and auditory perception. A user study with twenty-four children (Gaggi et al., 2012) proved that the games are feasible for preschool children and showed that the activities are engaging. Preliminary results in a follow-up study showed promising results in the link between task performance and a possible reading and spelling deficiency (Gaggi et al., 2017).

Regarding the English-speaking world, there are websites that offer a service for screening language skills with subsequent automated evaluation. Often, the exercises are set in a playful setting, such as the DORA Dyslexia Screener (Learn, 2023). However, most of the tools require a certain level of reading knowledge and focus mainly on detecting reading difficulties. The approach for pre-readers needs to be simple (e.g., tablet vs. desktop) and should not assume existing literacy or phonological awareness knowledge. To the best of our knowledge, there is no digital, game-based application that provides reliable screening of reading and spelling skills for pre-readers.

Table 2.2: Collection of digital and/or game-based literacy screenings.

Screening		Authors	Established/ Updated	Language	Test type	Medium	Game-based	Duration	N	Testing age	Focus
PB-LRS	Gruppentest zur Früherkennung von LRS	Barth and Gomm (2004b)	2004/2019	German	Group test	Analog	Gamified	60 min.	N=474	Preschool and first grade	Validity
RdH	Rundgang durch Hörhausen	Frank et al. (2001)	2002/2014	German	Individual test	Analog	Yes	45 min.	unknown	Preschool and first grade	Validity
HASE	Heidelberger Auditive Screening in der Einschuluntersuchung	Schöler and Brunner (2008)	2007/2008	German	Individual test	Analog and digital	No	10 min.	N=52832	Preschool (4-6)	Validity
MusVis	Screening with Musical and Visual Elements	Rauschenberger et al. (2022)	2018/2022	German/ Spanish	Individual test	Digital	Yes	10 min.	N=313	Pre-readers	Feasibility & Validity[a]
Dleksia	Dleksia Game	Mohtaram et al. (2017)	2017	Malaysian	Individual test	Digital	Yes	unknown	N=11	Primary school (6-8)	Feasibility
Gaggi Screening	Series of Serious Games to Screen Dyslexia	Gaggi et al. (2017)	2017	Italy	Individual test	Digital	Yes	unknown	N=24	End of Kinder-garten (5)	Feasibility
DORA	Series of Serious Games to Screen Dyslexia	Learn (2023)	2023	English	Individual online test	Digital	Yes	10-15 min.	unknown	Primary school (2. Grade)	Validity

[a]Rauschenberger et al. (2017), Rauschenberger et al. (2020a)

Although using digital, game and game-based screening to identify children with dyslexia before they can read and write sounds very attractive, it is essential to remember that these tools only provide trends and indications of dyslexia and cannot replace a medical or therapeutic diagnosis. Therapeutic interventions delivered in individual or group sessions are the most common approach to treating children with learning disorders. These interventions are provided in learning facilities by trained practitioners, such as teachers or learning therapists, usually outside school hours.

Chapter 3

Playful Testing

The presented diagnostic tool aims to screen children with weak literacy skills using a unique approach called playful testing. In this chapter, we try to define this new term by putting it into perspective to similar systems like serious games, stealth- and game-based assessment. In the last two decades, a lot of research and development regarding the aforementioned approaches promoted the use and assessment of educational digitization whilst accessing the advantages of a game environment.

3.1 Serious games

Plato already described games as a way to shape an adult when played during childhood (Ifenthaler et al., 2012). The perception of games by society has changed back and forth throughout the years. They have been viewed as entertainment for the masses and used as a distraction from politics in the Rome society (*panis et circenses* - bread and circuses (Bernstein and Bernstein, 1998)), to an illegal - even evil - activity due to their distraction from work (Dirx, 1981; Permentier, 2004), to a source of recreation from work (Kant and Rink, 1803). Only at the beginning of the 19th century did it become clearer that games can positively affect people, as its valuable effects were seen in children's development (Ganguin, 2010). The emergence of games into scientific territory in the 20th century was majorly supported by contributions from Freud (1920), Huizinga's *Homo Ludens* (Huizinga, 1955), Piaget (1975) and Dörner and Bick (1983). With the increased interest in games and integration of their advantages into non-entertainment-specific content, we have seen an enormous gain of publication in social science focusing

on so-called serious games (Ifenthaler et al., 2012).

Games whose primary purpose is to entertain, make fun or recreate are part of the digital commercial entertainment industry. However, serious games or game-based learning mainly aim to train and change the user's behavior. Entertainment or fun can still be part of the design but are more considered to be in a supportive role (Connolly et al., 2012). Serious games have widespread use in different areas such as education, research, advertising, health or politics (Djaouti et al., 2011) and can, therefore have different goals depending on their intended use. The game-based learning method often gets mixed up with the serious games approach. It is important to differentiate these two clearly. Game-based learning aims towards educational teaching and providing content to learn something in a game-based way (Connolly et al., 2012). Due to the fact that the goal of the presented diagnostic assessment tool is not to improve the player's performance on the tasks but to test the player's behavior and skill set in a playful way, we will continue to focus on serious games and game-based assessment for the rest of this dissertation.

3.1.1 Definition and use of serious games in an educational context

There have been many attempts and iterations to define serious games since it was first written down by Abt (1970). Djaouti et al. (2011) discuss in their meta-analysis "Classifying Serious Games" that there are generally two types of definitions. A general description of serious games and domain-specific definitions that account for differences across the industries in which the games are used. Although our presented literacy screening fits mainly into the educational domain, the broader definition of serious games can help us understand the intentions of the screening and how it can lead us to a better definition of our work. According to Djaouti et al. (2011), serious games are defined as "[...] any piece of software that merges a non-entertaining purpose (serious) with a video game structure (game)". A serious game consists of a serious dimension as well as the game dimension. A common misconception is that gamifying a task is enough to categorize it as a serious game. The structure relies on various components playing off of each other (e.g., storytelling, interface, reward- and redeem-system) and packaging it as a unit.

3.1.2 Advantages of educational games and serious games

Serious games have a lot of potential to bring individuals closer to a specific topic. Nowadays, audiences are more and more drawn to gaming, as the PC gaming industry shows a constant increase of its revenue by over 80% from 2014 to 2020 (PwC, 2020). This trend is expected to continue. The increasing use of digital devices in conjunction with games is a trend that can also be observed among the younger generation. With this demand in mind, the gaming and educational industry increasingly invests in serious games (Dondlinger, 2007). Especially in a sensitive context - like dyslexia - where the audience may find it challenging to deal with tasks and exercises, a playful approach to learning and teaching is beneficial. Although games are widespread across different industries - including education - according to Derryberry (2007), they all share some common attributes. To implement a successful and playful application for assessment, we must understand what constitutes a target-oriented game.

Backstory and storyline : The reason why a game is played, is mainly explained through the story and the storyline it follows.

Game mechanics: Game mechanics describe the world's physical behavior and the actions triggered by a specific input by a character.

Rules: Every player's actions and skills are limited due to the applied rules, which conclude from the game mechanics.

Immersive graphical environment: 2D/3D graphics, audio and animation help to represent the story in an appealing way to the player.

Interactivity: Interactivity is an essential feature of a game (Thornton and Cleveland, 1990). It shows the implication of a player's action and its impact on their surroundings.

Challenge/Competition: Challenges against oneself or any other character make the game experience exciting.

Risks/Consequences: Consequences are why the player is interested in taking on the aforementioned challenges.

With the help and combinations of these aspects, learners benefit from educational games in terms of their motivational, emotional, social and cognitive impact (Ge and Ifenthaler, 2017). Frustration and boredom can be successfully addressed, motivating the player

and keeping their attention to continue learning by playing the game (Deterding et al., 2011). These findings are supported by several empirical evidence (Boyle et al., 2016; Bressler and Bodzin, 2013; Clark et al., 2016; Eseryel et al., 2011) and were also confirmed by user evaluations as well as user tests of games with preschoolers (Barendregt, 2006; Hanna et al., 2004; Zaman, 2008). As Connolly et al. (2012) points out, the sum of these aspects can eventually lead to a positively shaped learning curve. As studies further show, mobile serious games can help children overcome their learning disorders independent of location and time. Mobile serious games specifically for dyslexic children have been proven to help in the process of literacy acquisition (Berkling and Pflaumer, 2014; Kast et al., 2011; Rello et al., 2014). A field study, which was conducted by members of the *Prosodiya*-Project at the University of Tübingen, shows the positive effect the serious game *Prosodiya* has in an interventional context (Holz et al., 2023). The digital game-based intervention engaged and motivated the participants over 14 weeks, significantly improving their writing ability.

It is important to remember that not everyone playing a serious game is familiar with its game mechanics or used to playing games at all. The audience can be very broad. It has to be applicable to all of the target audience and must therefore be designed as such. In the context of *Prosodiya* child-friendly, visual aesthetics were used as well as playful, pedagogical agents to engage with the children and facilitate learning (Holz et al., 2023). Difficulties in understanding the goal or usability issues can be frustrating and negatively affect their learning experience, even more so for inexperienced players (Moreno-Ger et al., 2012).

3.2 Assessment

To understand how games can support educational measurement and assessment of educational goals, we first need to take a look at the process of assessment and stealth assessment in general.

3.2.1 Assessment in general

Most children know from an early age that what they have learned at school will eventually be tested. Measuring the outcome of those tests - the degree to which knowledge, skills or attributes have been acquired through teaching - is called educational measurement (Shute and Ventura, 2013a). Tests can be used to collect and analyze data from

learners but are not designed to improve learning or the educational outcome (Snow and Jones, 2001). Assessments are designed to go one step further, interpreting the test information collected and responding to it. Furthermore, teachers, parents and eventually the students themselves can use the information in a facilitating way, for instance, to improve the test or test design. Therefore, measurement can only be seen as a necessary tool and a part of assessment (Shute and Ventura, 2013a).

3.2.2 Stealth assessment

Most classroom assessments are conducted as a separate event. Once the skills have been acquired in class, the evaluation of these skills follows. In this way, the educational instruction process is interrupted. According to Shute and Ventura (2013a), assessment should be a constant flow of information to facilitate student monitoring, as is common in retail stores, for example, where inventory stocks are monitored continuously rather than just a few times a year. In addition, assessment should be invisible, support in real-time and provide timely feedback (Shute and Levy, 2009). The term stealth assessment is generally associated with an assessment that is embedded into games to "[...] unobtrusively, accurately, and dynamically measure how players are progressing relative to targeted competencies." (Shute and Ventura, 2013a). This type of assessment sometimes provokes debate. In the advertising industry, stealth assessment can be misused without the user's knowledge to obtain information that would otherwise not be accessible. Despite these concerns, when implemented ethically, stealth assessment presents a powerful tool for personalized learning, offering real-time feedback and tailored educational experiences without disrupting the engagement or enjoyment of the game.

3.2.3 Game-based assessment

We now know that a game environment in a serious context can benefit children in terms of motivation and attention. We also know that stealth assessment is a way to continuously assess players' competencies without interrupting the game or the learning process. Ifenthaler et al. (2012) and Ifenthaler and Kim (2019) define game-based assessment in three parts: Game scoring, external assessment and embedded assessment. With scoring, the user acquires a target they aim at and knows what obstacles to overcome. External assessments are tests or (de-)briefing interviews that can be held during,

before or after the game and interrupt the game flow to explicitly measure what the user has learned. Embedded assessments are measurements that don't interrupt the game and typically take place in the backend of the software to log additional information about game performance and behavior.

In general, the game-based approach to assessments helps to reduce dropout and test anxiety while not sacrificing validity and reliability (Shute, 2014). However, in the context of a screening that aims to assess children's literacy deficits at a pre-reader age, we have to reevaluate the game-based assessment approach.

3.3 Serious games and game-based assessment in the context of literacy screenings for pre-reader

> "You learn more about a person in an hour of play than in a year of conversation."
>
> – *Plato*

For the present screening, which is intended to identify possible reading and spelling deficiencies in children at an early stage, the positive, playful elements are as important as the correct and successful implementation of the exercises. The playful elements familiar from serious games help us to build up the screening in a comfortable and motivating way. Children at a pre-reader age are not very familiar with exams or their skills being tested. Furthermore, the goal of a screening test is not to assess how well a specific content has been learned or practiced, as most serious games aim to do. In the context of stealth assessment, Shute and Ventura (2013b) refer to the fact that "[...] a test does not typically improve learning any more than a thermometer cures a fever; both are simply tools". In contrast to stealth assessment, children do not need any feedback or assistance when performing the screening tasks. The assessment should therefore be the main task of the screening. Since the assessment of children tends to take place during the game and is not intended to test what has been learned, the focus of a game-based screening is on embedded assessment rather than external assessment. As Shute and Ventura (2013a) point out, good games don't pre- and post-test players, as teachers might. Instead, the games seamlessly assess people as they play them.

We do not know of a definition for this unique form of assessment - with motivating and engaging elements from serious games, embedded stealth assessment, and yet goal-

oriented execution of the tasks. We therefore categorize the presented screening as a serious game with testing characteristics or, in short, as playful testing.

Part B

Screening Study

Chapter 4

Developing the Screening

In this section, we first examine the requirements of the screening to be feasible in group settings and discuss the game design elements, followed by the detailed explanation and design rationales of the five language-independent tasks included in the screening.

We focused on two main factors when we designed and developed the screening for young children, specifically in a pre-reader age: usability and the screening's feasibility in group settings. That is, the screening needs to be intuitive and easy to use and its tasks have to be easy to understand. Secondly, it needs to motivate and engage the children to continue playing and minimize distractions to make it feasible in a group test setting.

We programmed the screening software with the Unity[1] game development environment (version 2019.4.16f1) in C# and deployed it for the Android operating system.

4.1 Development process

Similar to the release cycle procedure known from software development (Dooley, 2011), we developed the screening software iteratively and repeatedly checked and tested with multiple pilot studies, see Figure 4.1. We included children during the design and development process as users, testers, informants, and design partners, as proposed by Druin (2002). This way, we avoid that design decisions are not adapted to the target group, which may pose barriers and make the screening less accessible.

[1] `https://unity.com/`

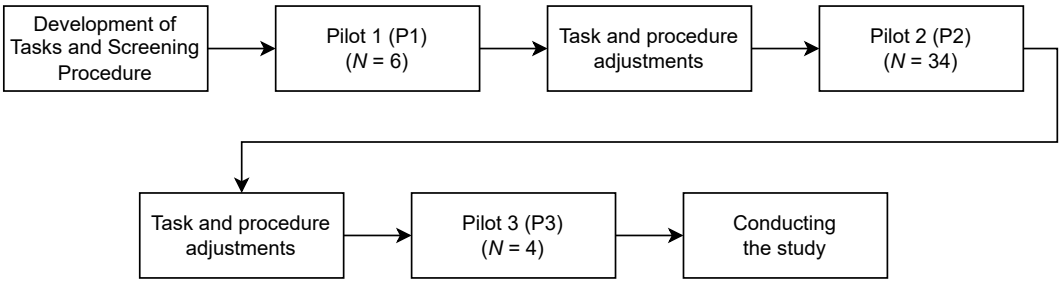


Figure 4.1: Several pilot studies were conducted between development stages to refine the screening process gradually.

The pilot studies conducted, their primary objectives, and the findings we were able to draw from them are listed in detail in Table 4.1. Conducting several pilot tests at an early stage to determine (i) the technical limitations and (ii) the social behavior of the children in the group proved essential in identifying and addressing potential risks at an early stage.

4.2 Individual testing in a group environment

4.2.1 Screening setup

We conducted an initial pilot test in an early stage of the development with $N = 6$ primary school children aged $8-12$ years to evaluate prerequisites to prevent distractions, suitable hardware (e.g., tablet stands, headsets, tablets) for the children and the general feasibility of the screening. Our observations indicate that it is important to isolate each child as much as possible to minimize visual and auditory distractions from and to others. The best results for reducing visual distractions were obtained by spatially dispersing the children across the available space (e.g., classroom) and assigning each child their own table. We reduced auditory distractions best with closed headphones, on which the volume could be adjusted separately. Tablet stands also proved beneficial for presenting the tablet at a comfortable angle to children. We have chosen tablets as a presentation device because they are easy to use, practical and usually familiar to children, as studies show (Radesky et al., 2020).

Table 4.1: Setup, Goals and Findings of the conducted pilot studies, leading up to the final version of the screening study. In addition to the technical implementation of the screening, it was also important to examine which prerequisites are necessary for a test environment that is as free of interference as possible.

	Pilot 1 (P1), $N = 6$	Pilot 2 (P2), $N = 34$	Pilot 3 (P3), $N = 4$
Setup	• 2nd - 4th grader • 3 groups of 1-3 children • completed each task in turn • individual and small group test in a non-school facility • qualitative and quantitative individual feedback after each task, recorded by hand	• 2nd - 3rd grader • 9 groups of 2-10 children • group tests in a partner school • first iteration of complete screening procedure with all intermediate steps • user- and game-experience questionnaires after each task	• 1st grader • 1 group of 4 children • group tests in a partner school • final iteration of complete screening procedure including a paper task for measuring the intelligence metric
Goals	• finding best setup to reduce distractions • identify major problems with the tasks based on the children's feedback	• feasibility of screening in a group test environment • check, that tasks are processed independently by the children • testing data logging	• testing the final screening setup with the target group • concretize the spoken instructions for the testing personnel
Findings	• best fitting hardware • optimal seating arrangement • added countdown at the beginning of the tasks	• more gamification elements needed • the task order that least affects the impact of fatigue	• ideally, at least two testers are present • minor bugfixes in data logging

4.2.2 Playing independently

We designed the application so that each child can independently complete the screening. The navigation within the application is designed so that only a straightforward procedure is possible and there is no back and forth. There is no possibility of making "wrong" decisions that change the sequence or to get lost while performing the tests, e.g., by skipping tasks. We achieved this by limiting the possibilities of user interaction between tasks. When used for testing, the application is in a state called *user mode*, which ensures a fixed task order. If the application crashes, the exam administrators can restart it in *admin mode*, allowing them to continue the test at the appropriate point.

4.3 Design of game elements

In the following, we describe the game elements used in the screening. The game elements keep children motivated to complete the screening independently and successfully. As mentioned in Section 3.1.2, they are essential to motivate and engage children, maintain focus and reduce disruptions as well as test anxiety. These elements proved indispensable and very helpful, especially when the screening was conducted in a group setting.

4.3.1 Sticker rewards

Rewards are often used as an integral game element in game-based training and game-based assessment to motivate the students (Deterding et al., 2011) and reduce dropout rates (Gaggi et al., 2017). In a group-based setting with multiple students taking the test simultaneously, rewards through an external force, such as the screening administrators or teaching staff, are not feasible and desirable, disrupting the assessment process. Therefore, a seamlessly integrated reward system is preferred. After completing each task, the children receive a colorful and child-friendly animal sticker as a reward, see Figure 4.2b. This sticker is representative of the task just finished. Their presentation is animated and accompanied by a splash sound, making them more appealing. A collection of all the stickers collected so far is shown when the children are about to start the next task. They are displayed on a wooden sign as an overview, as seen in Figure 4.2a, in which stickers for not yet completed tasks are indicated by placeholders. The stickers primarily serve as a motivating factor to continue playing. Still, in the overview, the children can also see how far they have progressed overall in the screening

(a) Virtual sticker collection. Children collect a sticker for each screening task.

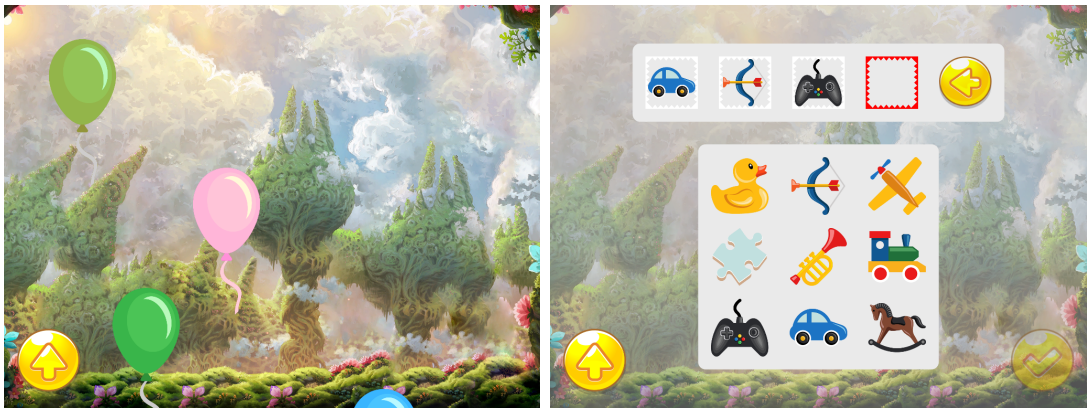
(b) An exemplary sticker that is rewarded to the children after they complete a screening task.

Figure 4.2: Virtual rewards used in the proposed screening to motivate and engage children.

and how many tasks are left. We had the impression that this reward system seemed to be fulfilling its task of motivating and engaging. Several children proudly talked about their rewards afterward and wanted to compare their achievements with others. Moreover, after receiving the rewards from each task, we perceived joyful reactions from the children.

4.3.2 Balloon minigame

In game theory, the principle of relaxation after tension has been well established (Kiili et al., 2012). Translated to our screening, this means that after a cognitively demanding task, it is advantageous to relax and not to further cognitively stress the children. The challenging task here is to develop a relaxation phase that does not lead to a state that provokes potential distraction. We implemented a simple minigame to keep the children engaged with the tablet even after finishing their current task. This minigame ensures that children who already completed the task do not disturb other children still engaged in a screening task. After the children finished a given task of the screening and received the respective sticker reward, the balloon minigame started. In this minigame, colorful balloons float across the screen from bottom to top that children can pop by tapping on them, see Figure 4.3a. Their spawn locations and rising speeds vary slightly. No other game elements, such as a reward or score system, are implemented in the minigame to



(a) Balloon game to keep the children engaged when they finish a screening task.

(b) Design elements used to ensure the screening's feasibility in groups.

Figure 4.3: Virtual rewards used in the proposed screening to motivate and engage children.

reduce the possibility of too much excitement and engagement. For the same reason, we present no auditory feedback or other objects. Only one button that opens the symbol keypad (see next section) to unlock the next task is visible and interactive. Our observations confirmed the importance of this task. Children who finished their tasks earlier than others were still engaged with the tablet and the balloon minigame. Overall, we observed far fewer interruptions due to boredom or overexcitement after the children finished a task than in previous pilot iterations without this game element.

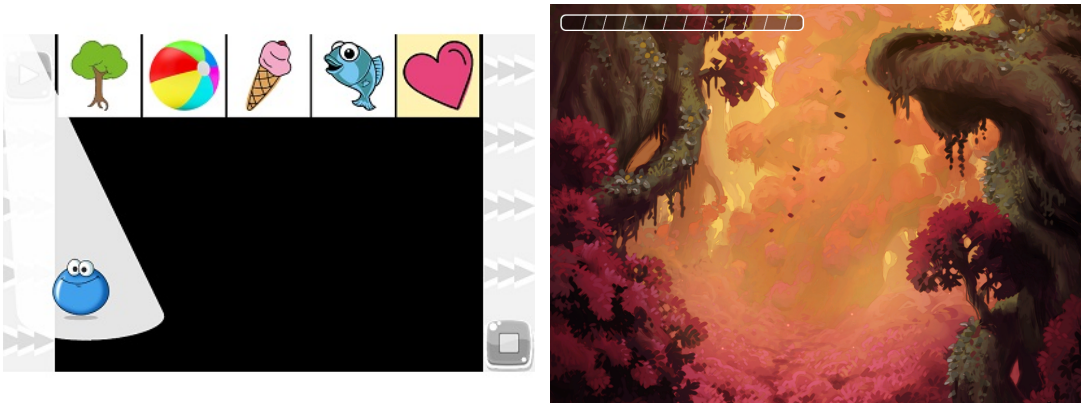
4.3.3 Symbol keypad

In a group setting, different skill levels coincide. While some children find it easy to complete a task, others may find it more challenging and need more time. If this circumstance is left unregulated, a child may not be able, for instance, to focus on the agent's position in the SRTT because a neighboring child is naming the objects of the RAN task aloud into their headset. To reduce potential distractions and confusion caused by asynchronous task completion, it is advisable to ensure that children begin each task simultaneously. External software can be used to achieve this by remotely controlling the display of each tablet or the application itself. However, after further research, it turned out that this approach is associated with too many unknown risks (e.g., technical and spatial requirements at the test site, technical know-how of the instructors, unknown

time management due to longer setup times) and cannot be implemented without further investment of resources (e.g., purchase price, licenses, personnel). A simple way to prevent children from starting the next task ahead of time is a lock system that requires a code to unlock the next task. Each task - except the first one - is unlocked with a specific four-digit combination of symbols. This combination is only known to the instructor. When instructed, all children open up the symbol keypad from the balloon minigame screen after each task. Nine symbols are displayed in a 3×3 grid with four empty fields above the grid, as seen in Figure 4.3b. When a child clicks on an object from the grid, the symbol of the object is placed in the next empty position in the upper field. Children can click the *check* button to evaluate the combination once all positions are occupied. If the symbolic combination is correct, the next task is unlocked. In our study, printing out and showing bigger versions of the symbols on paper was beneficial when reciting them to the children at the beginning of the screening. This way, we ensured every child knew what symbol to click on and bypass any possible language barriers or uncertainties on what to do. Children perceived the symbol keyboard as part of the screening along with the aforementioned balloon minigame. Our observations indicate that including these elements did not deviate the children from the screening and its game flow, as might be the case if the screening was paper-based or not game-based.

4.3.4 Interactive and self-explanatory tutorials

In traditional, individual (paper-based) screenings, instructors mostly need to repeat the task instructions for each person. Digital instructions with interactive elements can help to improve and standardize this process. The screening needs to have clear and child-friendly instructions to ensure that children can understand the process and carry it out independently, even in a group test (Chen, 2017). Players view tutorials as essential phases in onboarding a new game, where they play a crucial role in aiding their understanding and successful completion of the game (Zichermann and Cunningham, 2011). Each task presented in the screening has its own structure and game mechanics and therefore requires its instructions beforehand. When starting a new task, children must always complete the tutorial first. In our case, pedagogical agents in colorful and small spherical lights accompany and guide the children in each tutorial, see an example in Figure 4.4a. They help explain game mechanics to the children and act as companions during the screening. Companions are often used in this context to engage children (Lim and Reeves, 2010). The interactive design of the tutorials can increase



(a) Exemplary tutorial for the RAN task. The blue agent explains the task and its mechanics.

(b) One example of a fantasy-themed environment with the goal to motivate and engage children.

Figure 4.4: Interactive tutorials and fantasy-themed environment.

children's participation and ensure their understanding of the game mechanics. After agents instruct or guide the children on what to do, they are tasked with helping the agent solve multiple test trials during the tutorial. After each tutorial, we must ensure that it covers the entire scope of an actual task trial. Children at a young age have a limited attention span and cognitive load. To address this, it is important that tutorials are short and straightforward but do not cognitively overwhelm children with too much new information in a short period. The presence of the agents and the interactivity helped to address this problem. We revised the tutorials in multiple iterations based on observations and feedback from pilot tests with children to optimize the extent of the tutorial and the balance between length and complexity.

4.3.5 Visual aesthetic

We embedded the Screening in a fantasy-themed environment. Research shows that such environments positively impact motivation, engagement and learning (Cordova and Lepper, 1996; Parker and Lepper, 1992). For the proposed screening, we adapted the environment of our previously successfully evaluated spelling intervention for primary school children (cf. Holz et al., 2023). With this environment, we could embed the user interface and game elements into an inherently coherent environment. Using different backgrounds for each task gives them a unique setting. We designed the companions with fun and appealing behaviors towards the children. We also created all stimuli for

the tasks in a distinctive and child-friendly manner.

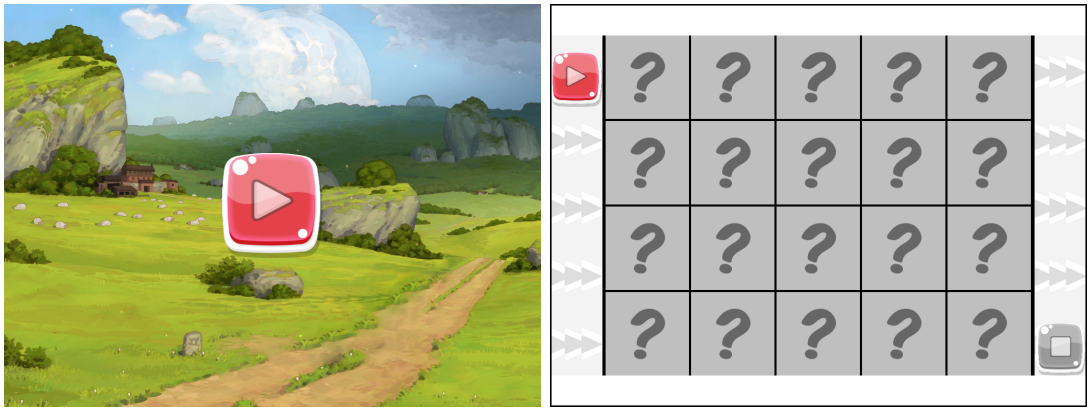
4.3.6 Further game elements

As mentioned, game elements increase motivation, engagement and enjoyment in a learning context (Huotari and Hamari, 2012). Additional game elements were important to further resemble the character of the screening to that of a game.

Progress and feedback Each task's progress, except for the SRTT and RAN tasks, is displayed as a progress bar at the top of the screen, indicating how far the children have progressed. The progress bar updates after each trial by filling out the next segment. In addition, neutral auditory feedback is played to further indicate the completion of a trial. For the two tasks that do not feature a progress bar, the insight into the progression is either not desired (SRTT) or redundant due to the design of the task (RAN).

Traffic light system Each child can set their own pace of play and decide when to begin tasks. They have to actively press the green *play* button to start the tutorial of the next task, as seen in Figure 4.6(b). The primary outcome variable for the SRTT and RAN tasks is children's reaction time, i.e., naming time and clicking on the stimulus, respectively. Therefore, drawing the children's attention to the task is important. A traffic light indicator accomplishes this by changing its signal from red to yellow to green once the child presses it, accompanied by an audible countdown sound at each stage to inform the child when the task starts. The children then mentally and physically prepare and are ready for the start of the task. An example of this traffic light indicator can be seen in Figure 4.5a and Figure 4.5b.

Navigation We implemented a straightforward, one-way navigation through the tasks to prevent children from bypassing or repeating unintended steps. For this, only one clickable element at maximum is displayed on each screen to allow navigation. An example can be seen in Figure 4.6 that shows the sequence of the RAN task as an exemplary sequence for other tasks in the screening.



(a) Traffic light indicator before the SRTT.

(b) Traffic light indicator within the RAN task.

Figure 4.5: Traffic light indicators used to draw children's attention and to prepare them for the beginning of the task.

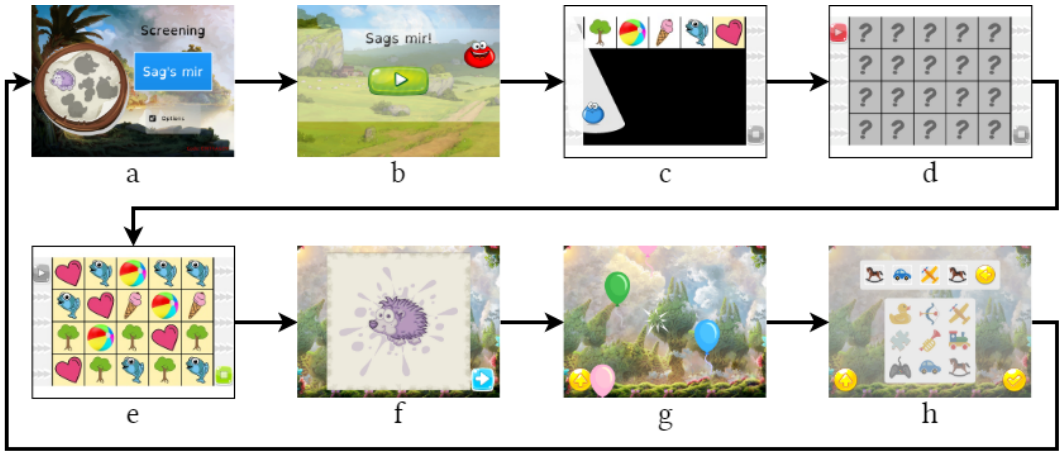


Figure 4.6: Sequence of the RAN task as an example for the general task procedure. Starting from the main menu (a), the children have to start the tutorial by pressing the green *play* button (b). After completing the tutorial (d) and main task (e), the children receive a reward (f) and get forwarded to a minigame (g) while waiting for the rest of the group to finish. When all children have finished the current task, they enter the four-digit symbolic code (h) told by the instructor to unlock the next screening task. When the children enter the correct code, the main menu pops up again to start the next task (a).

4.4 Tasks

In the following sections, we give a brief overview of the theoretical background for each task, the creation process for the task items and stimuli, and describe the task process in detail. As we have established from the game and user experience study, it is of great advantage to children if the tasks are presented digitally and in a playful manner. In this chapter, we address these aspects of the implementation of the tasks, among others, and thus answer *RQ1.3*.

4.4.1 Incidental Holistic Perception Task (IPET)

4.4.1.1 Overview

There is evidence that reading difficulties, as they occur in the context of developmental dyslexia, are associated with the holistic perception of visuospatial information (Von Károlyi et al., 2003) and enhanced recognition of incidentally encoded, speech-free stimulus material (Hedenius et al., 2013). The Incidental Holistic Perception Task (IPET) is a recognition task focusing on incidental holistic perception by combining these two approaches in a novel and digital setup. In this task, children are asked if they have seen specific images in a previously briefly presented grid. Based on the design of the study of Hedenius et al. (2013), these images are either real animals or made-up fantasy creatures. After the encoding phase, eight images are displayed one after another and the children are asked for each image whether they have seen this image (*yes*) in the previously shown grid or not (*no*). See Figure 4.8 for a depiction of the IPET procedure.

4.4.1.2 Background

Researchers have mainly associated dyslexia with various cognitive, linguistic, or non-linguistic deficits. Over the years, more and more research has been conducted on potential improvements in specific cognitive functions (Lovegrove et al., 1982), for example, showed that under certain conditions, boys with reading difficulties had advantages over control subjects in the processing of visuospatial information. This finding was later supported by the adoption of Stein (2001), who predicted that the weakness in the visual magnocellular system[2] could lead to an improvement of the parvocellular system[3].

[2]Neurons in the retina, which are primarily responsible for the perception of movement and contours.
[3]Neurons of the retina, which are responsible for color perception, spatial resolution, pattern and detail analysis.

Stein (2001) assumes that this could result in more efficient, holistic processing and more accurate storage of large images, settings or events as shown in different studies (Lovegrove et al., 1982; Tafti et al., 2009). Although the strength-oriented approach is rising, it is noteworthy that the findings in this area are relatively ambiguous and inconsistent.

The finding that people with reading difficulties process visuospatial information more efficiently when the task requires global holistic processing of that information nevertheless appears to be the most robust. Both Von Károlyi et al. (2003) and Diehl et al. (2014) show that young people with reading difficulties react more quickly than control subjects in the presentation of seemingly impossible figures[4] and are at least as accurate. The aforementioned authors were able to replicate similar results from a previous study by von Károlyi (2001).

More recent results also provide evidence for such processing advantages in connection with memory tasks. In their study, Hedenius et al. (2013) were interested in the declarative memory performance of dyslexic and typically developed children. The object recognition task conducted by Hedenius et al. (2013) was split into three parts. Encoding (marking the presented stimuli as *real* or *made-up*), recognition after 10 minutes, and recognition after 24 hours (recalling if they have seen a shown image in the encoding phase with *yes* or *no*). The 11 children with developmental dyslexia (DD) had better accuracy when recalling seen images compared to their 17 typical developed (TD) peers, as shown by their one-way ANCOVA with accuracy as dependent variable and performance IQ as covariate ($M_{DD} = 86.4\%,\ SD_{DD} = 13.9\%;\ M_{TD} = 73,9\%,\ SD_{TD} = 20.3\%,\ p = .075$). The two groups did not differ in their reaction times. As a result of their findings Hedenius et al. (2013) suggest that the performance disadvantages that occur in the context of declarative memory performance may not be the result of inadequate memory performance. It mainly happens when those people are asked for uncommon verbal content, explicitly requested to remember such content, and when asked to reproduce it actively. They assume that the shortcomings in those memory performances might arise from dysfunctional phonological information (remembering verbal content) processing and executive dysfunctions (producing verbal content) rather than insufficient memory abilities.

[4] Optical illusions that cannot exist in a 3D space

4.4.1.3 Task construction

Based on the work of Hedenius et al. (2013), our main premise for the task is to make distinctions according to the incidental recognizability of real and made-up objects. We decided to use real and fictional animals to convert the structures and constructs used in the original study into child-friendly objects. We show commonly known animals in a cartoon-like but real appearance. In contrast, the fictitious animals are cartoon-like creatures that might look like dragons, little monsters, or some variations of them. We took care to avoid objects that might scare children. We also added the holistic approach of Von Károlyi et al. (2003) by displaying 12 animals simultaneously in one grid image. Considering the time constraints of the overall screening as well as the time query of recognition in the study of Hedenius et al. (2013) (recognition task after 10 minutes and after 24 hours) and our goal to integrate the approaches of incidental encoding with holistic perception, we had to develop a new task design, which can be seen in Fig. 4.8.

We piloted different approaches for the grid image display. Adding a corresponding fictitious or realistic environment for the presented animals to give them more context, such as the background of a farm or an outer space area, did not seem to add any value and only made the evaluation of the stimuli more complex. We also found that many of the results of the pilot study were at the level of guessing. To minimize the possibility that this effect occurs due to the too-short display duration, we increased the display time of the grid image from 500ms to 700ms for the main study. To exclude any bias regarding the position of the objects, we evenly distributed them over the grid image instead of randomly scattering them.

Since we did not find a differentiated evaluation of the animals in the pilot study, as we only showed either real or fictitious animals in each grid, we decided on a grid of six real and six fictitious animals and distributed them randomly across the grid. For each grid, we randomly selected two objects from both animal types for later recognition and marked them as *target*. Additionally, we made sure to query two objects from each position in the grid as target objects. We allocated two real and two fictional objects to each grid image, which were not part of the original grid image but were selected as *distractors* for the recognition task. We ensured that each object was used only once in a complete run-through. We generated a total of two sets, each containing six grid images. In the first set, the objects that served as targets became distractors in the second set,

and vice versa. This ensures the removal of any bias related to type association (target vs. distractor). When testing, each group was assigned one of these two sets before the start of the test so that both sets were distributed equally among all groups. We bought all 96 image objects needed for the task (1 set * (12 objects + 4 distractors) * 6 run-throughs) from the image database Adobe Stock[5] with a similar style of art.

4.4.1.4 Task procedure

The task begins with a tutorial where the red and blue agent explains the gameplay and guides the child through some training trials. See Figure 4.7a for a screenshot of the tutorial and appendix Chapter 12 for a translated transcription. A trial consists of the presentation of a grid image composed of 12 stimuli, some of which are later to be distinguished from distractors by the child in a recognition phase. One grid image contains 12 stimuli that are ordered in a 4x3 grid. In order to direct the child's attention to the center of the screen, a fixation cross is presented for 1000ms beforehand. After showing the centered grid image for 700ms, a grey masking image is shown for 700ms to overwrite the working memory. Then, eight objects are displayed one after another and the child is asked after each object whether he thinks this object was present in the grid image (procedure can be seen in Figure 4.8). In total, there are six grid images and 6x8 recognition trials. After a response has been registered by clicking the green or red button, a short acoustic signal and an update of the progress bar at the top of the screen indicate the end of a successful trial. The feedback is always the same and independent of the given response.

4.4.2 Syllable Stress Task (SST)

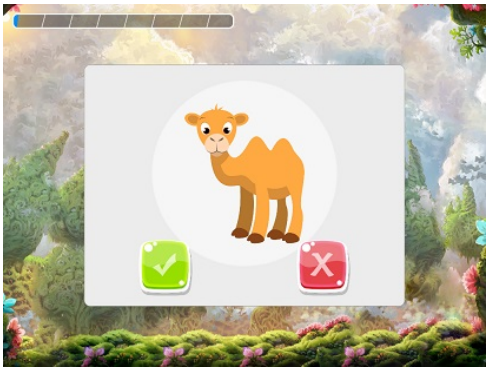
4.4.2.1 Overview

The Syllable Stress Task (SST) is based on the empirical evidence that the ability to recognize (syllable) stress is impaired in developmental dyslexics (Goswami et al., 2013; Jiménez-Fernández et al., 2015; Leong et al., 2011) and that it correlates highly with reading and spelling skills (Brandelik, 2014; Sauter et al., 2012). The underlying cause seems to be a deficit in auditory perception (Huss et al., 2011). Children must identify the appropriate stress pattern for target sentences in this task. To do this, the sentence is first read aloud and paired with an illustration of the sentence. Then, two different stress

[5] `https://stock.adobe.com/`

(a) The red and the blue agents explain the procedure of the task through training trials in the tutorial.

(b) A grid image with 12 different objects is shown at the beginning of the test trial for a very short time.



(c) The query of recognizing eight different items is checked after the initial grid image.

Figure 4.7: In the IPET, the children have to correctly recognize or reject an object after being presented with a grid image with multiple objects for a short time.

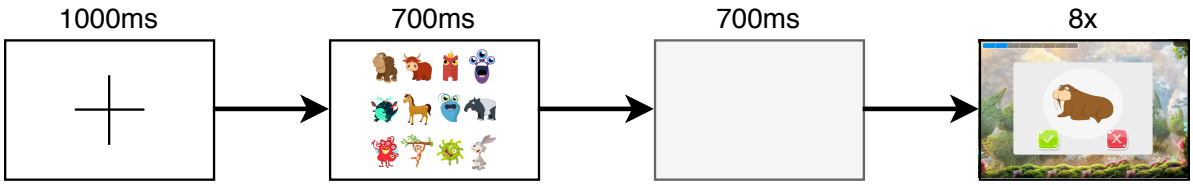


Figure 4.8: Sequence of a single grid image followed by eight recognition trials with a masking image in between.

patterns, one of which matches the stress pattern of the target sentence, are presented visually and played on a piano. The children must choose the stress pattern that they think matches the rhythm of the sentence. The procedure can be seen in Figure 4.9 and Figure 4.10.

4.4.2.2 Background

The lack of phonological awareness is one major cause of dyslexia (Bradley and Bryant, 1983). Phonological awareness also includes the perception of prosodic features, in which shortcomings strongly predict developmental dyslexia (Goswami et al., 2013; Leong et al., 2011; Sauter et al., 2012). Syllable stress is part of these features and, due to German being a stress-timed language (Barry, 2003), is considered an important aspect of German speech rhythm. Alongside this, auditory perception plays an additional role in recognizing prosodic features. Identifying auditory structures in language and working with them is crucial for learning to read (Huss et al., 2011). Meter in music is equivalent to syllable stress in language. Studies show that auditory processes are used in music and speech to extract structure and rhythm from both (Corriveau and Goswami, 2009). As part of this structure, metrical perception is important to separate words and syllables from a speech stream (Hura and Echols, 1996). The syllable stress task presented builds on the fact that the ability to recognize syllable stress is impaired in children with developmental dyslexia (Goswami et al., 2013; Jiménez-Fernández et al., 2015; Leong et al., 2011). Further, Sauter et al. (2012) and Brandelik (2014) found high correlations between writing skills and a speech-based stress task ($r = .78,\ p < .0001$), and between voice length errors and accuracy in recognizing stress patterns ($r = -.68,\ p < .0001$). Therefore, the language-based stress task used in these studies is the basis for the SST.

4.4.2.3 Task construction

The children have to complete sixteen trials, four of which were practice trials. The first two practice trials were presented with only one response stress pattern to get to know the test in general and its procedure. A complete transcription of the tutorial can be found in appendix A.2.1. After the practice trials, the children could continue whenever they felt ready for the main trials by clicking on the green start button.

With the help of a linguistics expert, we created the sentences for each trial with the intention of escalating their difficulty as the task advanced. Simple and short sentences with few syllables, few vowel length markers and no compound words become more complex in word length, structure and stress pattern over time. We have ensured that the sentences' complexity is appropriate for the target audience to prevent a lack of concentration and distraction. We recorded the sentences at a rate of 2-4 syllables per second with the help of a professional speaker[6]. One sentence contained a compound word whose main accent was marked as a stressed syllable and the secondary accent was marked as an unstressed syllable ('Schlitten,fahren - Dddd).

Regarding the audio, we normalized the volume of all response patterns to -3 dB relative to the maximum sound pressure level to avoid clipping. However, the loudness of the final presented stimulus depended on how high each child set the volume of the tablet device. In a response pattern, a big dot results in a long and loud piano note (∼500ms-550ms), whereas a small dot results in a short and quieter piano note (∼450ms) with a frequency equal to a G4 at ∼391 Hz. Besides the sentences, the distractor response patterns also have different difficulty levels that increase as the task progresses. The difficulty of a distractor is defined by its distance in prosodic features to the corresponding target response stress pattern. A greater distance is based on more differences in these features, making the distractor less difficult. We categorize the distractors into three levels of difficulty, ranked from 1 (easy) to 3 (hard). In the first category, the distractors differ in stress and number of syllables. Items from the second category contain at least three different stress strokes, whereas items from the third category contain two at most. A list of all sentences and response patterns used in this task alongside the different difficulty categories for distractor patterns with the extension of Hamming and Levenshtein distances can be found in the appendix Table A.1.

[6]Comparison: the normal speaking rate in German is on average 4.45 syllables per second (Gebhard, 2012)

(a) The yellow agent leads the tutorial, which explains the process and game mechanics, like selecting a stress pattern.

(b) After the sentence and both stress patterns were auditive and visually presented, the child has to decide on one stress pattern.

Figure 4.9: In the SST, the children have to match one of two presented stress patterns to a sentence that is read aloud and visually supported by an image at the beginning of the task.

4.4.2.4 Task procedure

Children are presented with an auditory and visual stimulus at the beginning of the task. A read-aloud sentence is given and supported by an image in the middle of the screen. After the sentence has been read out once, the image is scaled down and two response patterns (target and distractor) are presented successively below the image. A response pattern consists of black dots that are representations of a stressed (big dot) or unstressed (small dot) syllable. The response pattern is visually presented while simultaneously being played as a tonal pattern. Only one of the two response patterns corresponds to the stress pattern of the given sentence (target). The stressed beats and the corresponding larger dots can be synchronized with the stressed syllables of the sentence, while the unstressed beats and the corresponding smaller dots can be synchronized with the unstressed syllables. The distractor rhythm can not be synchronized with the stress pattern of the sentence. The children's task is to choose a response pattern that matches the stress pattern of the sentence. After both patterns have been displayed, the children can listen to each component again and answer by clicking on one of the two patterns and confirming their input by pressing the continue button on the bottom right. The next sentence is presented after a short universal, auditory feedback independent of the given answer. Screenshots of the task can be seen in Figure 4.9.

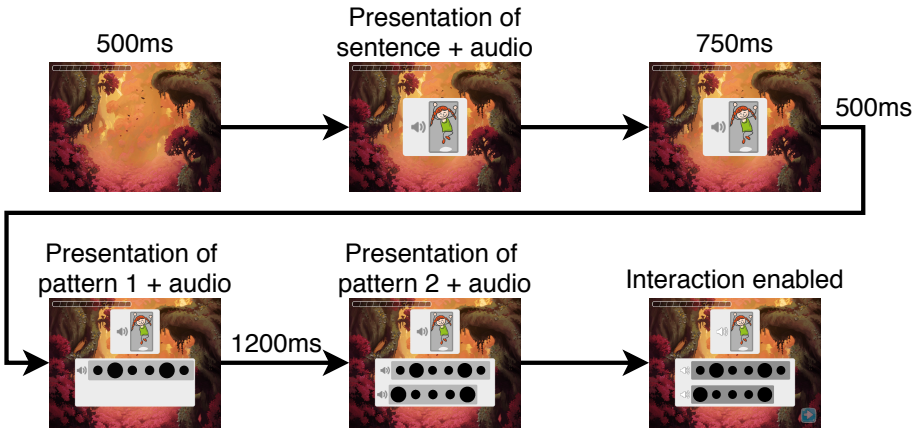


Figure 4.10: SST procedure with time specifications for stimuli and audio presentation.

4.4.3 Rise Time Discrimination Task (RTDT)

4.4.3.1 Overview

In addition to the auditory limitations in recognizing stress patterns, the deficit in recognition and differentiation of rise times in sound seems to distinguish dyslexic children from their control peers (Huss et al., 2011). Therefore, with a discrimination test, the Rise Time Discrimination Task (RTDT) further explores auditory perception - specifically in rise time. In the RTDT, children are asked to find the tone that matches a previously presented target stimulus based on auditory properties (rise time and steady state). The RTDT incorporates simple sine waves as language-independent stimuli, as suggested by Rauschenberger et al. (2017), with specific rise time, steady state, and fall time. The target sound differs in either rise time or rise time and steady state to the distractor sound. A screenshot of a task trial can be seen in Figure 4.13b.

4.4.3.2 Background

People with dyslexia have difficulties in auditory and visual perception (Ortiz et al., 2014). Those shortcomings seem to be due to issues with the short-term memory (Overy, 2000). Most of the screenings described in Section 2.2 focus on using visual perception to learn or assess. To explore auditory perception even without the need for prior linguistic or prosodic knowledge as in the Syllable Stress Task, we can use music and some of its relevant features as a reliable predictor (Rauschenberger et al., 2017).

Impairment in basic auditory processing affects language and musical development. Both domains evolve in time and therefore have some sort of rhythm (Huss et al., 2011). Es-

pecially speech-language impaired (SLI) children show difficulties in auditory cue processing in the language domain (Corriveau et al., 2007; Huss et al., 2011). Rise time and rhythm seem to be critical auditory cues for timing in music and language. A series of studies investigated auditory perception (e.g., perception of rise time) of SLI children. They showed that their rhythm perception is impaired (Corriveau and Goswami, 2009; Thomson and Goswami, 2008) while their pitch recognition simultaneously is intact as quoted by Huss et al. (2011): "[They have] brains that are in tune but out of time.".

The most significant advantage of using a rise time perception metric is that no direct linguistic knowledge in any form is necessary to perform the task (Rauschenberger et al., 2017). The presented task displays a combination of Rauschenberger's web-based game approach to detect dyslexia with elements of sound as well as integrate an established predictor in using rise time discrimination from the studies conducted by Huss et al. (2011). Hämäläinen et al. (2013) showed in their meta-analysis about basic auditory processing deficits in dyslexia that all investigated studies regarding discrimination in amplitude modulation and rise time displayed a significant correlation to developmental dyslexia. For our approach, we took a closer look at the study conducted by Huss et al. (2011). The study consisted of assignments on intelligence, phonological awareness, phonological short-term memory, perception of musical meter and six psychoacoustic tasks. The stimuli for the RTDT were constructed based on two of these six psychoacoustic tasks. In the Amplitude Envelope Onset (Rise Time) Task, the authors presented an AxB format for rise time discrimination. On each trial, the 8-11-year-old children were presented with three sounds, from which two were constructed using the same pitch, the same rise time, the same steady state and the same fixed fall time. The third sound differed in rise time but with the same total length of 800ms as the other sounds, making the steady state also different from the other sounds. The children's task was to decide which sound was different from the others. With a fixed total length of exactly 800ms and a changing rise time, the steady state would always vary between 450ms and 735ms. Therefore, it is possible that the children could discriminate the rise time stimuli not only due to the change in rise time alone but also to the change in steady-state duration. To rule out this cue, Huss et al. (2011) included a follow-up task, Rise Duration Rove Task, with the same setup as the Amplitude Envelope Onset Task but a randomly changing steady-state duration across the task. As a result, the length of the task also became variable. The results of the one-way ANOVA analysis

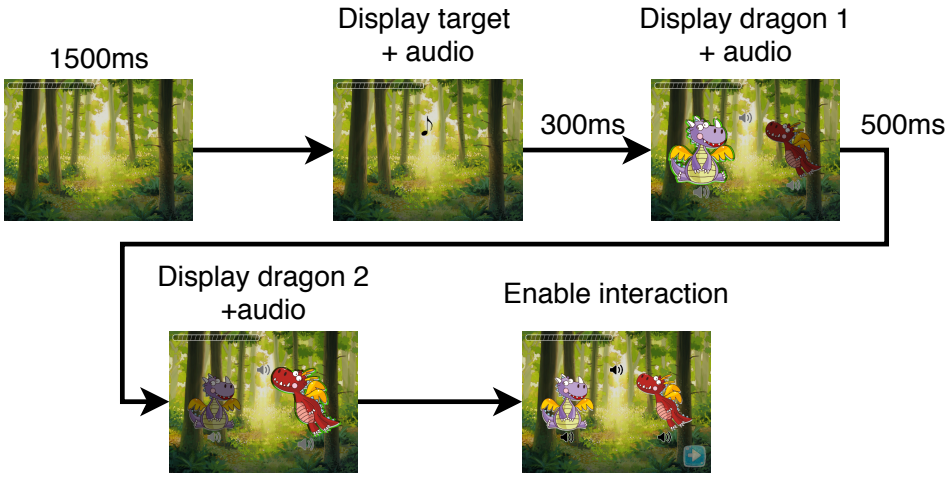


Figure 4.11: RTDT procedure with time specifications for stimuli and audio presentation.

conducted by the authors suggest that children with developmental dyslexia (DD) show lower performances as well as higher variance in both the Amplitude Envelope Onset Task ($M_{DD} = 105.9$ ms, $SD_{DD} = 73.0$ ms; $M_{TD} = 36.4$ ms, $SD_{TD} = 14.8$ ms, $p < .01$) and Rise Duration Rove Task ($M_{DD} = 113.4$ ms, $SD_{DD} = 76.3.0$ ms; $M_{TD} = 31.8$ ms, $SD_{TD} = 11.7$ ms, $p < .01$) compared to their age-matched control peers (TD). Although the Rise Duration Rove Task was aimed to improve rise time stimuli discrimination, it also introduces an additional bias. Stimuli with short rise time and random steady-state duration will generally be shorter than their counterparts with a longer rise time and the same random steady-state. In creating our rise time stimuli, we tried to find an in-between solution.

4.4.3.3 Task construction

We created 20 stimuli using the specification of Huss et al. (2011) and the technical methods used by Rauschenberger et al. (2017). We used the free software Audacity[7] to create all the musical stimuli. These stimuli consist of a simple sine curve at approximately ∼269 Hz (C4). When creating the sounds, we normalized the volume to -4 dB so that they can be played at 0 dB at maximum volume without overmodulation. Table A.2 in the appendix shows all used audio pairs and their acoustic features. We incorporated the Rise Duration Rove Tasks from Huss et al. (2011) described earlier to minimize the possibility of rise time discriminating other than by rise time, which aims to tackle the

[7] https://www.audacityteam.org/

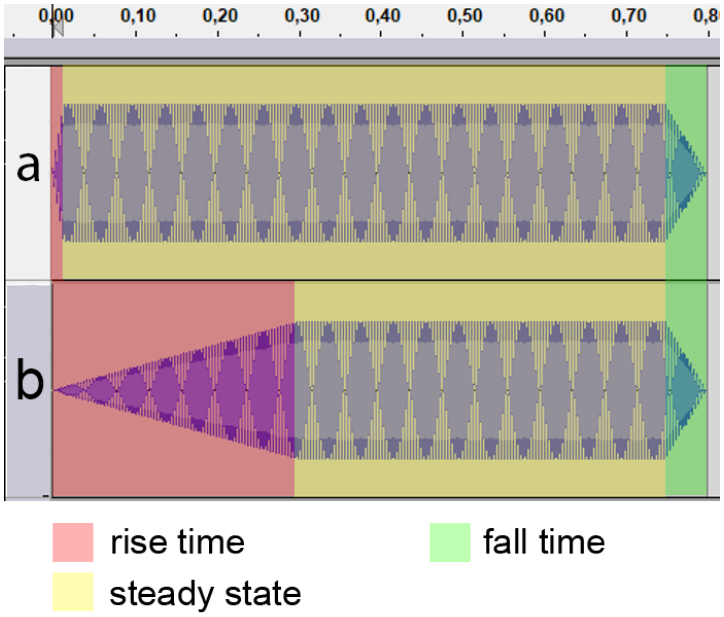
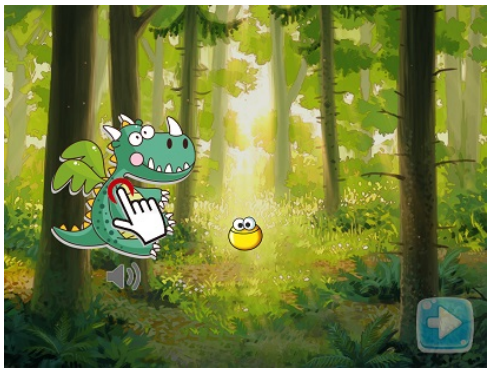


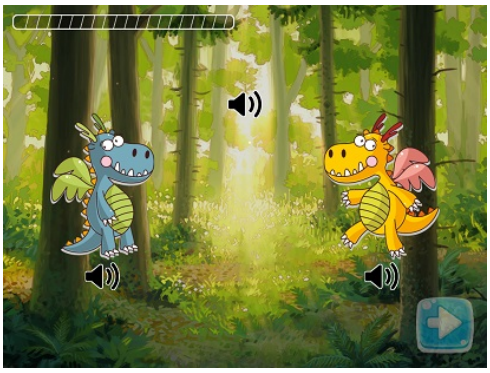
Figure 4.12: Track *a* shows the structure of an item with 15ms rise time and track *b* shows the structure of an item with 300ms rise time, both from category 1.

discrimination problem due to a variable steady-state length. We therefore created two item categories. Category 1 describes items oriented to the Amplitude Envelope Onset Task with variable rise time between 15ms and 300ms and an adaptive steady-state duration, which results in a fixed total length of 800ms. Items from the second category, which are oriented to the Rise Duration Rove Task, have a variable rise time between 15ms and 300ms and a pairwise, fixed steady state duration between 450ms and 735ms, resulting in a variable total length of the sound. All sounds have the same fall time of 50ms. An exemplary structure of the items can be seen in Figure 4.12.

As far as we understand, the Rise Duration Rove Task from Huss et al. (2011) was primarily used as a control instance for the Amplitude Envelope Onset Task and therefore wasn't mainly used to measure the differences in rise time perception. As a result, our final item set consists of item pairs from both categories in the ratio 3:2 (category 1:category 2) and was randomized once for all children. The full set of all stimuli used in this task can be found in the appendix Table A.2.



(a) In the tutorial, the yellow agent introduces the task and explains how to select and deselect dragons.



(b) The child has to decide which dragon made the same sound as the one presented at the beginning of the task.

Figure 4.13: In the RTDT, the participants must identify the dragon that makes the same sound as the reference tone. Figure 4.13a shows the yellow agent explaining the task in the tutorial and Figure 4.13b displays a test trial when waiting for input.

4.4.3.4 Task procedure

This task is set in a forest environment. The task is explained to the child by letting them complete two training tasks together with the yellow agent at the beginning of the task. A detailed tutorial description can be found in the appendix A.3.1. A trial begins with an auditory stimulus represented by a symbol of three eighth notes animated in the back of the forest. The children are instructed beforehand to memorize this reference tone. They are then presented with two different-looking dragons that enter the scene. Upon arriving and starting with the left dragon, they successively make a sound, which is either the same as the reference sound or slightly different in rise time alone or in rise time and length. Until both dragons present their sound, they are not interactable. Before deciding by clicking on one of the two dragons, the children can listen manually to all the sounds presented. After selecting one dragon and confirming the child's input by pressing the button to continue the task, a sound effect indicates the end of the trial.

4.4.4 Serial Reaction Time Task (SRTT)

4.4.4.1 Overview

The Serial Reaction Time Task (SRTT) has established itself over the years as a test for measuring implicit and procedural learning (Frensch and Rünger, 2003; Nissen and Bullemer, 1987; Robertson, 2007). As other research further suggests, it can be used as a predictor for dyslexia (Lum et al., 2010). In the SRTT, children are asked to respond quickly to stimuli at different screen positions. Unnoticed by the children, the stimuli follow a repeated sequence and do not appear at seemingly random positions. After several blocks of repeated sequences, a block of random sequences is presented. While reaction times improve in the blocks with repeated sequences, they slow down in the blocks with random sequences, suggesting an implicit sequential learning effect. Research also suggests that children with literacy weaknesses generally react more slowly in this task (van der Kleij et al., 2019). A trial procedure within the SRTT can be seen in Figure 4.14b.

4.4.4.2 Background

In 1987, Nissen and Bullemer introduced the Serial Reaction Time Task (SRTT), which allowed them to measure implicit spatial learning (Nissen and Bullemer, 1987). In their study, a light stimulus appeared in one of four places. Participants had to press one of four buttons that corresponded to the location of the stimulus. The participants were divided into two groups. One group was unknowingly given a sequence of ten consecutive light positions, which was repeated ten times. This resulted in 100 trials per block, and a total of eight blocks were performed. In the other group, the positions of the light stimuli were set randomly after each trial so that subconsciously, no sequence effect was recognizable. The only restriction in the random condition was that no consecutive trials could have the same position. Plots show that the reaction time decreases considerably over time when the position of the stimuli follows a fixed sequence. In contrast, a random sequence of the position does not lead to a noticeable improvement in reaction times. The two-way analysis of variance furthermore confirms significant main effects of group (repeated vs random) as a between-subject factor with ($F(1, 22) = 53.17,\ p < .0001$) and block as a within-subject factor ($F(7, 154) = 25.86\ p < .001$) as well as a significant interaction between group and block ($F(7, 154) = 11.93\ p < .001$) (Nissen and Bullemer, 1987).

The SRTT is a standard method to demonstrate implicit and procedural learning (Frensch and Rünger, 2003; Robertson, 2007). There is no clear definition for the term implicit learning. However, many researchers define it as learning without knowing what one has just learned (Frensch and Rünger, 2003). For example, a person can walk without being able to describe the processes that take place in their body while walking. Acquiring motor skills and behaviors by performing tasks are the basis of procedural learning (Koziol and Budding, 2012).

A broad field of studies suggests a connection between disorders in language acquisition and impairment of implicit learning processes. Research by Vicari et al. (2005) showed that children and adolescents with a diagnosed reading and spelling disorder were impaired in various implicit learning tasks, including the SRTT. In a study by Stoodley et al. (2008), only children with reading and spelling disabilities showed no implicit learning in the SRTT, both in comparison with peers who are at least in the normal range and in comparison with peers who are reading and spelling impaired but do not meet the double discrepancy criterion. Lum et al. (2010) also find significant differences to the disadvantage of speech-language impaired (SLI) children on the SRTT when comparing the reaction time differences of both groups (SLI and not-SLI) between the last training block and the pseudorandomized block. The authors suggest that there is a deficit in procedural memory, as this is responsible for implicit learning. This learning effect has also been demonstrated in studies where the SRTT was performed in digital form on a tablet (Minuth, 2019; Moisello et al., 2009). These studies have shown that the difference in reaction times on the digital medium is often more than half that in traditional SRTT studies. Further research shows, however, that the difference in reaction times between the training blocks and the randomized block can be the same for both groups but that the children with dyslexia are overall significantly slower in their reaction time when recognizing objects (van der Kleij et al., 2019).

In 2013, Lum et al. (2013) published a meta-analysis of 14 studies comparing the learning performance in the SRTT of people with and without reading and spelling disabilities. On average, the children with reading and spelling disorders performed about half a standard deviation worse in the SRTT than those in the control group. Lum et al. (2013) found that control groups generally show more sequence learning in the form of a larger reaction time difference between sequence and random block than dyslexic groups. The difference decreases when more training to sequence is given (for example, by adding more training blocks) and participants are older. They also point out that

some of the item sequences used in the studies vary between 5 (Vicari et al., 2003) and 12 items (Deroost et al., 2010; Rüsseler et al., 2006), indicating that the differences may only be visible in longer sequences due to easier learning with fewer items (Howard and Howard, 1989).

4.4.4.3 Task construction

The structure of the SRTT is based on the study by Lum et al. (2010). The four positions are arranged in a diamond shape in the center of the screen. There are a total of five blocks, consisting of four consecutive training blocks followed by a block with pseudorandomized sequences. The ten-item sequence is repeated five times in each training block, resulting in 50 trials per training block. The sequence is identical to the original sequence of Nissen and Bullemer (1987) and was also used by Lum et al. (2010): 4 2 3 1 3 2 4 3 2 1 - respectively: SOUTH NORTH EAST WEST EAST NORTH SOUTH EAST NORTH WEST. Assignments to the positions can be seen in Figure 4.14c.

According to the findings of Lum et al. (2013), the longer the sequence, the more meaningful it tends to be and avoids possible learning effects within a sequence. The fifth block is pseudorandomized but derived from the original ten-item long sequence following two constraints. First, the displayed stimulus appears on the platform the same number of times as in training blocks 1-4. Second, the probability of the stimulus appearing on a platform, considering the position of the preceding platform, was kept the same as in training blocks 1-4. Table 4.2 shows the constraints to create the pseudorandomized sequence in the fifth block. These restrictions in randomization exclude pair association learning as a cause for differences between the sequence blocks and the random block.

Considering the conclusions that Lum et al. (2013) drew from their meta-analysis, the compact structure presented and the early testing in school should lead to a better differentiation between the groups. In contrast to the study of Lum et al. (2010), the task was performed on tablets that allowed the child to touch the stimuli directly on the screen rather than using an external controller[8] or buttons on the screen. This change was made after an early pilot version, which had buttons on the side of the screen (depending on the child's handedness) to be used with a thumb as input options, showed relatively

[8]We did not test external controller input as an option, due to its complexity in implementation and limited resources within the project.

Table 4.2: Constraints for creating the pseudorandomized sequence in the last block of the SRTT, derived from the 10-item long original sequence. The resulting 50 items long, pseudorandomized sequence can be found in the appendix A.4.1.

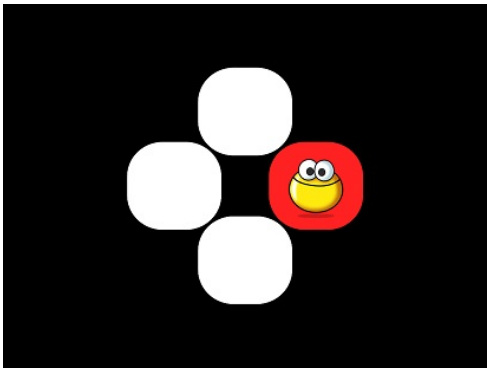
Position	Occurrence	Positions afterwards	Probability
NORTH	30%	{EAST; SOUTH; WEST}	{33%; 33%; 33%}
SOUTH	20%	{NORTH; EAST}	{50%; 50%}
WEST	20%	{EAST; SOUTH}	{50%; 50%}
EAST	30%	{WEST; NORTH; NORTH}	{33%; 33%; 33%}

slow reaction times due to higher complexity in eye-to-hand coordination. Further pilot experiments suggested that the direct touch of the stimulus is more intuitive and pleasant for the children.

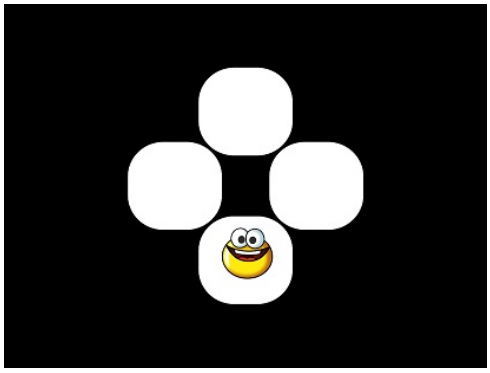
4.4.4.4 Task procedure

This task is split up into five blocks. The premise of each block is to 'catch' the yellow agent by clicking on it as fast as possible. The agent can appear on one of the four platforms, which are positioned in all four compass directions (North, South, East, West) in the center of the screen. Once a click is successfully registered on any empty or occupied platform, the agent changes its position onto another platform. The children were instructed to react as quickly as possible without making mistakes. The presence or absence of a repeating sequence was not mentioned. There are 50 trials to complete in each block. A small recovery break is implemented between blocks, in which the children have to explicitly press a button when they are ready to continue with the next block. After pressing the button to proceed to the next block and before starting the task, a short fixation cross is displayed for 1000ms to focus the child's eyes on the center of the screen.

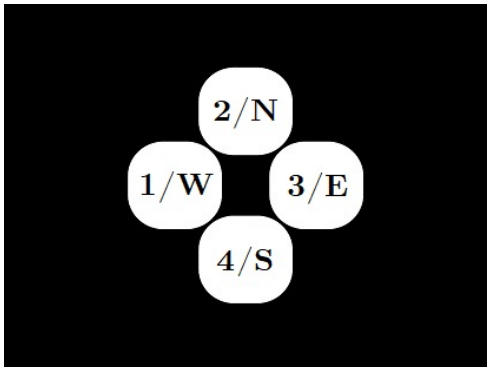
Instructions are given by the yellow agent at the start of the task. There are two interactive trainings set for the children to complete. A detailed transcription of the tutorial can be viewed in the appendix A.4.2.



(a) Extract from the tutorial, in which the yellow agent guides the child and highlights the platform on which the child is supposed to click next.



(b) This is an example from a test trial, in which the yellow agent jumped to the SOUTH position.



(c) Each position is assigned a label. W=WEST, N=NOURTH, E=EAST, S=SOUTH. To compare with the original sequence of Nissen and Bullemer (1987), numbers are also added.

Figure 4.14: In the SRTT, children have to click on the yellow agent, which then changes its position repeatedly according to an underlying hidden sequence. Figure 4.14a shows the tutorial, in which the children get introduced to the task. Figure 4.14b shows a test trial and Figure 4.14c displays the task structure and position naming.

4.4.5 Rapid Automatized Naming Task (RAN)

4.4.5.1 Overview

Naming speed was a major predictor for reading speed and therefore associated with reading in general (Wimmer, 1993). The Rapid Automatized Naming (RAN) task is a well-known tool to measure this speed indicator (Araújo et al., 2015; Denckla and Rudel, 1976; Moll et al., 2009). In the RAN task, children must name the stimuli on the screen aloud into their headset as quickly as possible. The RAN task measures both naming speed and naming accuracy. The children must name 40 items across two pages where they are presented in a four-by-five grid on each page. An example of a task trial can be seen in Figure 4.16c.

4.4.5.2 Background

In 1973, Denckla and Rudel (1976) found a way to differentiate dyslexic children by naming speed. Back then, the developed task required rapid repetitive naming of pictured objects, colors, letters and numbers. The authors described the decisive deficit as an automated verbal response and therefore called the task Rapid Automatized Naming. Since then, the concept of the RAN task has been used in a variety of studies to investigate its relationship to literacy further and the reasons for it.

As previously mentioned, phonological awareness is essential in predicting possible reading and spelling difficulties. Several studies on the effects of measures to improve phonological awareness show that mainly on reading, these effects are relatively small and only of short-term nature (Galuschka et al., 2014; Ise et al., 2012; Wolf et al., 2016). When considering reading skills independently, there are more meaningful variables to look at. Wimmer (1993) showed in a study with Austrian third graders that reading speed is the parameter that best differentiates between low and average-reading children. Brizzolara et al. (2006) similarly concluded that children with reading difficulties are characterized by massive difficulties with rapid naming, which they measured with the RAN task. At the same time, problems in phonological awareness were found mainly in the group of children with reading difficulties and additional language acquisition disorders[9]. The findings can be confirmed for real words and pseudo words, which test the applicability of the alphabetical principle (Araújo et al., 2015; Moll et al., 2009). In contrast to nam-

[9]The study was conducted in the Italian language, which is orthographically similar to the German language.

ing speed, reading accuracy could not solely be attributed to reading in children (Mayer, 2008; Wolf et al., 2002). This is probably due to the fact that in these studies, reading accuracy was realized solely by the number of reading errors without taking reading time into account.

In most of the mentioned studies, some variation of the RAN task was used. It measures naming speed - more precisely, lexical access and retrieval of information - and is a crucial predictor of reading performance. Deficits in naming speed is therefore a risk factor for children with developmental dyslexia (Araújo and Faísca, 2019). RAN is assumed to be closely associated with naming speed for real and pseudo words while reading accuracy and spelling performance are more closely correlated with phonological abilities, especially phonological awareness (Endlich et al., 2019; Moll et al., 2009). The performance in a RAN task often helps to explain individual variances in literacy skills over phonological awareness (Parrila et al., 2004).

Although the RAN task has a simplified structure, there is much discussion in the literature about what it measures and why success corresponds with literacy achievement. On the one hand, some authors try to explain the relationship between RAN and literacy skills in terms of their phonological similarity, e.g., efficiently retrieving phonological information from long-term memory (e.g., Wagner and Torgesen, 1987). This view is supported by research, for example, from Lervåg and Hulme (2009), suggesting that RAN measures the ability to identify and name objects, which is considered necessary for developing visual word recognition. On the other hand, some see the link between RAN and literacy either in a generally rapid information processing (e.g., Catts et al., 2002) or in the precise temporal coordination of information from different modalities (e.g., Wolf et al., 2000). These, in turn, are necessary for successfully merging phonological and visual information into orthographic skills to quickly recognize and process familiar and common symbols (Bowers, 1995; Powell et al., 2007).

4.4.5.3 Task construction

In contrast to the original structure of the task by Denckla and Rudel (1976), we had to limit the RAN task in its scope. On the one hand, to be able to complete the task in the given time frame and, on the other hand, not to overtax the children cognitively. The children were presented with two pages of five different but common objects placed randomly in a 5x4 grid. Pilot tests showed that with more than 20 stimuli on one page,

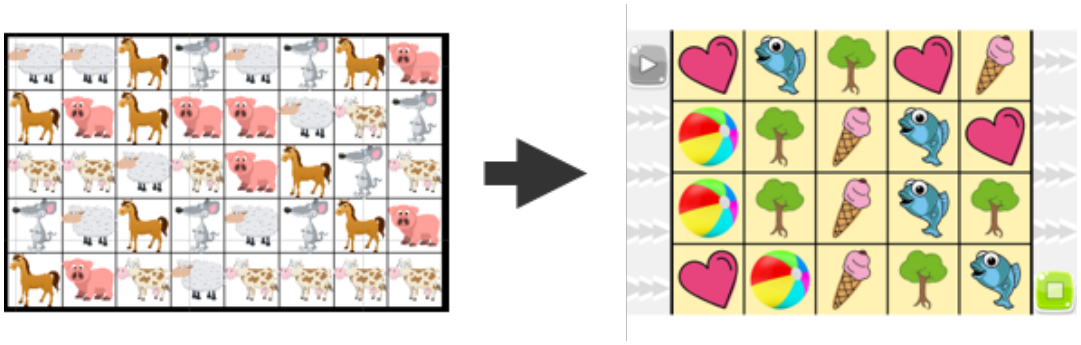


Figure 4.15: The first iteration of the RAN task consisted of 40 items on one page (left image), which was used in an early pilot test. The right image shows the second iteration with 20 items per page used in the final study.

each object's visibility and recognizability decreases, as seen in Figure 4.15. To still have as many stimuli as in comparable studies, we added a second page. Like previous studies, the children were instructed to follow the reading order from left to right and row by row. Visually, this instruction was emphasized by the arrows in the background pointing from left to right.

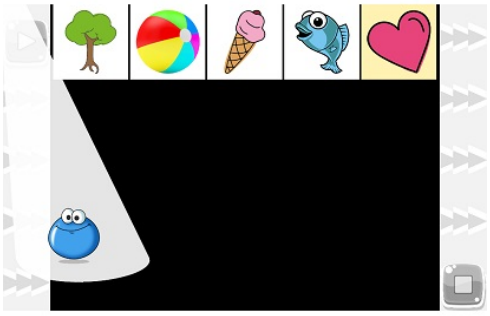
Five different objects were selected for presentation: heart (Herz), fish (Fisch), tree (Baum), ice cream (Eis) and ball (Ball). When choosing the items, it was vital that they are monosyllabic in German, highly-known[10] and child-friendly in appearance. Matching pictures were purchased on the image database Adobe Stock.

One major difference to previous studies was that the children had to mark the trial as finished by pressing the stop button. This is due to the overall goal to perform the task as a group test with no additional help from a supervisor. Consequently, a possible delay between naming the last object and pressing the stop button must be checked manually during a later analysis. The same applies to the delay between pressing the start button and naming the first stimulus.

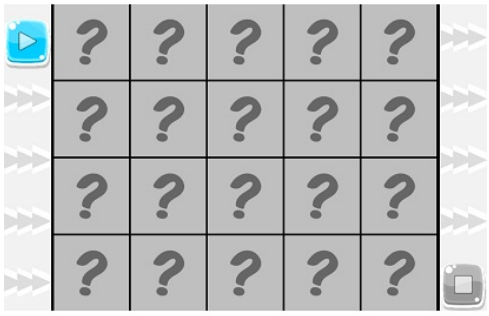
4.4.5.4 Task procedure

There are 20 face-down cards randomly arranged in a 5x4 grid on the screen. Once the experiment begins, the cards are revealed and the children are asked to read what

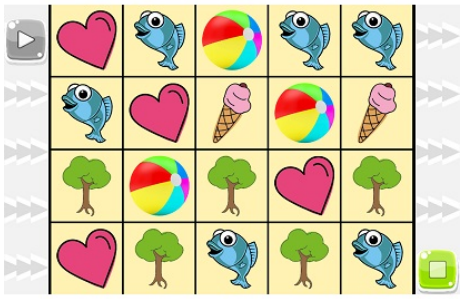
[10] All used words have a frequency class of 10 and below according to the corpus used in (Collection, 2018)



(a) The blue agents guide the child through the tutorial and present all possible symbols.



(b) Before a trial starts, all cards are hidden and faced upside down until the child presses the start button.



(c) After the start signal is given, all cards are revealed and the child is supposed to start naming all symbols from top left to bottom right.

Figure 4.16: In the RAN, children are asked to read different symbols aloud into their headsets. The blue agent introduces the children to the task and gives instructions on completing it, as seen in Figure 4.16a. Figure 4.16b shows the setup just before a trial starts. Figure 4.16c displays the distribution of the symbols after the child pressed the start button and a countdown finished.

they see on each card as quickly as possible from top left to bottom right into the headset they are wearing. The button on the top left corner with a play icon marks the starting point and must be pressed by the child to start the trial. The trial begins with an auditive and visual traffic light countdown. Afterward, all cards are revealed at once. The button with a stop icon in the lower right corner marks the end of a trial and must be pressed by the child as soon as the last symbol has been read. There are two pages to complete the task, accumulating to 40 stimuli in total. In the tutorial, a pedagogical agent first introduces all symbols that occur by reading them aloud and indicates to the child afterward that they get shuffled and placed upside down on the screen. Two trial pages follow, where on the first page, the agent takes over the trial and reads the symbols in the correct order while the input is disabled. On the second page, the agent retracts and lets the child perform the trial, including starting and stopping it. A detailed transcription of the tutorial can be viewed in the appendix A.5.1.

Chapter 5

Evaluating Game and User Experience

Parts of the following section are based on a manuscript we previously published (Holz*, Beuttler* et al., 2024).

This chapter takes a closer look at the second pilot test study (*P2*). In this study, we conducted and analyzed the game and user experience of the literacy screening. For this, we defined two overarching goals:

Q1. Examine the feasibility of the screening in a group setting, specifically answering the following questions:

  Q1.1. Can the proposed screening be used by young children independently without thorough instructions from adults?

  Q1.2. Are children engaged with elements of the screenings and do not distract other children?

Q2. How do (primary school) children - with both poor and (above) average reading and spelling skills - perceive the screening, i.e., what is their user experience and game experience, and do they consider the screening more as a test/exam or as a game?

The following sections describe the study and its findings and conclude how children perceived the screening. The R scripts and datasets supporting the findings of the game and user experience study are openly available and hosted on the Open Science

Framework[11].

5.1 Participants

Thirty-four German primary school children (17 girls) from second and third grade, aged $8 - 10$ years ($M = 9.00$, $SD = 0.74$), from seven primary schools in the area of Tübingen, Germany, participated in the pilot study. Twenty-five children from six partner schools participated in the study within the scope of a randomized controlled field trial of a game-based spelling training (Holz et al., 2023), of whom nineteen had poor reading and spelling skills. To increase the number of children with average and above-average reading and spelling skills, we additionally recruited eleven children from a partner school in the area of Balingen, Germany. The participants' demographics are listed in Table 5.1.

Due to technical complications, the questionnaire data of one child[12] is missing for the SRTT and for another child[13] for the IPET and SST.

Table 5.1: Demographic data of the participants.

	Typically Developing ($N = 15$)	Weak Readers/Spellers ($N = 19$)	**All** ($N = 34$)
Boys/Girls	7/8	10/9	17/17
Grade 2/Grade 3	4/11	7/12	11/23
Age [M (SD)]	9.07 (0.80)	8.95 (0.71)	9 (0.74)

5.2 Materials

5.2.1 Tablets and headphones

We used Samsung Galaxy Tab A 2016 tablets (type SM-T580) with 10.1-inch screen size and a resolution of 1920×1200 pixels running Android 7.1 in the study. We used tablet stands from Ikea (type ISBERGET) to position the tablets at a 45° viewing angle and connected the devices to NUBWO N2 headsets with suspension-style headbands

[11] `https://osf.io/hrbdv/?view_only=336b53066cfc4057b23043be750a2fb9`
[12] female second grader with poor reading and spelling skills
[13] male second grader with above-average reading and spelling skills

offering flexible and comfortable hold. To prevent the children from accidentally muting the speakers or microphones, we fixed the integrated volume and microphone controls with adhesive tapes.

5.2.2 Screening experience

We evaluated children's experiences with the screening with a questionnaire consisting of 29 questions that included 12 questions from the User Experience Questionnaire (UEQ) (Schrepp et al., 2017b), 13 questions from the Kids Game Experience Questionnaire (KidsGEQ) (Poels et al., 2008), and four self-designed questions to assess whether children experienced the screening more like an exam or more like a game.

5.2.2.1 User Experience Questionnaire (UEQ)

To assess children's user experience of the screening tasks, we used a subset of 12 items from the German UEQ for children and adolescents (Hinderks et al., 2012). This is a version of the regular UEQ but uses simpler language. The UEQ consists of bipolar terms that form a pair of opposites in the form of 7-point semantic differentials, e.g., **boring** ○○○○○○○ **exciting**, see Figure 5.1a. The items intended to measure the subscales *Attractiveness* (3 items), *Dependability* (1 item), *Perspicuity* (4 questions), and *Simulation* (4 items). The UEQ provides five benchmark categories for each subscale that relate to the observed mean scale values (cf. Schrepp et al., 2017a): *Excellent*, *Good*, *Above average*, *Below average*, and *Bad*. As we only used one item of the subscale *Dependability*, i.e., *hard to use/easy to use*, we renamed it to *Ease of Use*. The items were displayed on the tablet with up to four per page, see Figure 5.1a. The items and their Cronbach's alpha scale consistencies are listed in Table 5.2.

5.2.2.2 Game Experience Questionnaire (GEQ)

To evaluate children's game experience of the screening tasks, we used a subset of 13 questions from the Kids Game Experience Questionnaire (KidsGEQ) (Poels et al., 2008) that intended to measure the subscales *Positive Affect* (2 questions), *Competence* (3 questions), *Immersion* (2 questions), *Flow* (1 question), *Negative Affect* (3 questions), and *Tension/Annoyance* (2 questions). We translated the questions into German and adjusted them to relate to an exercise rather than a game. We used a 5-point word and color-coded rating scale with the following answer options: *not at all*, *slightly*, *moderately*,

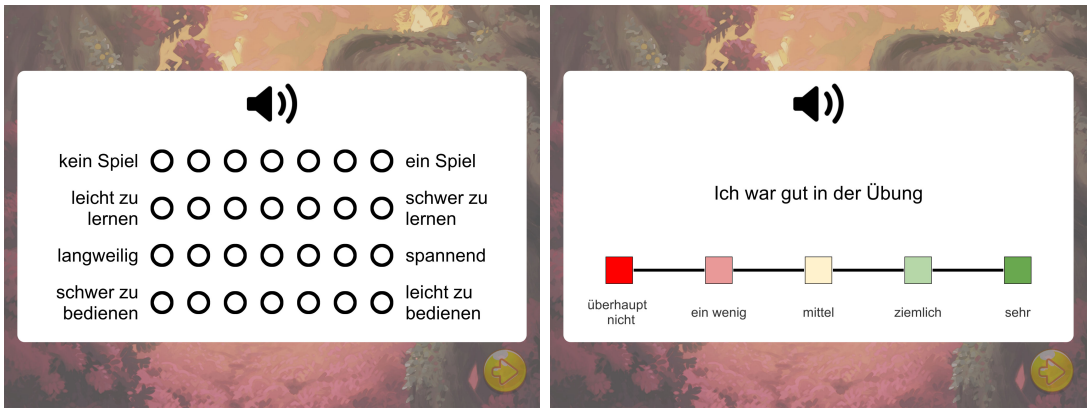
Table 5.2: Items used from the UEQ and KidsGEQ to measure user and game experience. Cronbach's alpha (α) is reported for scale consistencies. The items from the UEQ can be viewed in detail in the appendix Table A.3.

Subscale			
User Experience Questionnaire (UEQ)	*Items*[a]	α^{*b}	α
Attractiveness	[12, 16, 24]	.89	.66
Dependability / Ease of Use	[11]	.82	–
Perspicuity	[2, 4, 13, 21]	.82	.70
Simulation	[5, 6, 7, 18]	.76	.71
Kids Game Experience Questionnaire (KidsGEQ)	*Items*[c]		α
Positive Affect	• The exercise was fun for me • I felt good while working on the exercise		.57
Competence	• I felt confident during the exercise • I was good at the exercise • I felt I was up to the exercise		.71
Immersion	• The exercise was beautiful • The exercise was impressive		.68
Flow	• I was strongly focused on the exercise		–
Negative Affect	• The exercise was stupid • The exercise was tiresome/exhausting • I got bored with the exercise		.61
Tension/Annoyance	• I felt uncomfortable during the exercise • I was annoyed during the exercise		.62
Test Situation Questionnaire (TSQ)	*Items*[c]		
	The task/exercise was ...		
TSQ1	**no game** ○○○○○○○ **a game**		
TSQ2	**no test** ○○○○○○○ **a test**		
TSQ3	• I was afraid during the exercise that I would do something wrong		
TSQ4	• I felt like I was taking an exam during the exercise		

[a] Item number as reported in Hinderks et al. (2012).

[*b] Cronbach's alpha of the original UEQ as reported in Laugwitz et al. (2008).

[c] Translated from German.



(a) Exemplary items of the User Experience Questionnaire (UEQ) (Hinderks et al., 2012).

(b) Exemplary item of the Kids Game Experience Questionnaire (KidsGEQ) (Poels et al., 2008).

Figure 5.1: Visualization of the screening experience questionnaire.

fairly, and *extremely*, see Figure 5.1b. Each question was displayed individually and read aloud by the tablet. The children could repeat the questions by clicking the speaker *sound* button. The questions and scale consistencies are listed in Table 5.2.

5.2.2.3 Test Situation Questionnaire (TSQ)

To investigate whether the children perceived the screening tasks more as games or as exams/tests, we designed a *Test Situation Questionnaire* (TSQ) that contained a total of four items, see Table 5.2. The first two items used a similar bipolar rating scale as the one used for the UEQ and contained the pairs of opposites *game/no game* (TSQ1) and *test/no test* (TSQ2). The remaining two items use the 5-point word and color-coded rating scale of the KidsGEQ and assess whether the children were afraid of doing something wrong during the exercise (TSQ3) and whether they felt like they were taking an exam/test during the exercise (TSQ4). We integrated all items from the TSQ at random positions in the respective questionnaire with their appropriate scales: TSQ1 and TSQ2 into UEQ, TSQ3 and TSQ4 into KidsGEQ.

5.3 Procedure

We administered the screening sessions in classrooms of the partner schools during the lesson time in small groups of 2–10 children (average group size: $M = 4.86$, $SD = 3.94$). The test administrators prepared the tablets by entering the participant codes

and checking the volume level of the tablets. The children were seated individually at a desk with the tablets attached to the tablet stands and connected to the headsets. Before the screening started, we made sure each child was comfortable with the tablet and headset and that the headset's audio was working. In addition, the test administrators explained the rating scales used in the screening experience questionnaires in detail and provided explicit examples with mock-up questions.

We initiated the screening by instructing the children to simultaneously click the start button on their screen, which triggered the screening to start with the IPET task. Children first worked on the interactive tutorial before the actual task. After they completed the task, they were rewarded with a virtual sticker for their virtual sticker book. The procedure of a screening task is also depicted in Figure 4.6.

After they received the virtual reward, i.e., after each task, children were asked to answer the screening experience questionnaire for the respective screening task. For this, the children first answered the items of the UEQ, including two items of the TSQ. Four items were displayed on one virtual page and children were instructed to rate each item before they could proceed to the next page. The instruction, i.e., the sentence "What do you think of this task? How was it?" (in German: "Wie war die Übung für dich?") read aloud by the tablet, could be repeated by pressing the button with the sound speaker icon. After they answered the twelve items of the UEQ and the first two items of the TSQ, the children were asked to answer the thirteen items of the KidsGEQ and the last two items of the TSQ. Based on others' experience of children's unfamiliarity with such rating scales (cf. Holz et al., 2018), each question was displayed individually and read aloud by the tablet. At any point in the questionnaire, the children could repeat the audio for the question.

After the children completed the screening experience questionnaire, the balloon minigame kept them engaged with their tablets. The screening administrator waited until all children had finished the screening task and the corresponding questionnaire. The screening administrator then instructed the children to open the symbol keypad by pressing the yellow arrow button in the lower left corner. After all children successfully opened the symbol keypad, the administrator revealed the four-digit symbol combination to unlock the next screening task.

The remaining four screening tasks were carried out the same way in the following order: (2) SST, (3) RTDT, (4) SRTT, and (5) RAN. After all children finished the screening

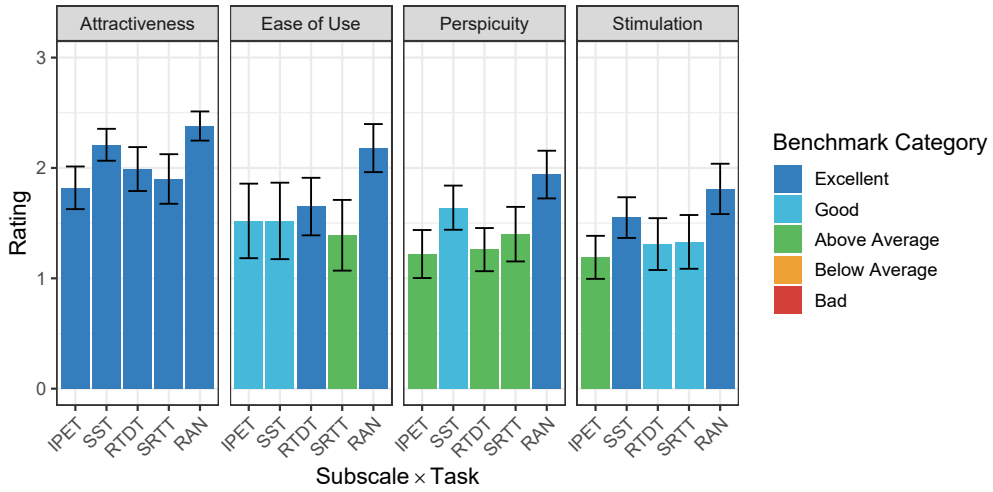


Figure 5.2: Children's user experience ratings by subscale and screening task (facet). The color indicates the benchmark category as classified by Schrepp et al. (2017a). Bars represent the standard error of the mean.

on the tablet by answering the screening experience questionnaire of the last task (the RAN task), they were rewarded with small toys and let go. The screening session took approximately 45-50 minutes.

The local ethics committee approved the study for psychological research and the Supervisory School Authority of Tübingen.

5.4 Results

We transformed the answers of the 7-point bipolar rating scales into values -3 to +3 (cf. Laugwitz et al., 2008; Schrepp, 2018) and the answers of the 5-point word- and color-coded rating scales into the values 1 to 5.

To contrast the influence of grade (second vs. third) and literacy skills (typically developing vs. poor readers/spellers) on the perceived screening experiences, we computed two-sample Wilcoxon rank-sum tests.

5.4.1 User experience

To evaluate the children's user experience, we computed the benchmark category for each subscale and task based on the classification of Schrepp et al. (2017a). Descriptive results, as well as the classification, are displayed in Figure 5.2. As can be seen in Fig-

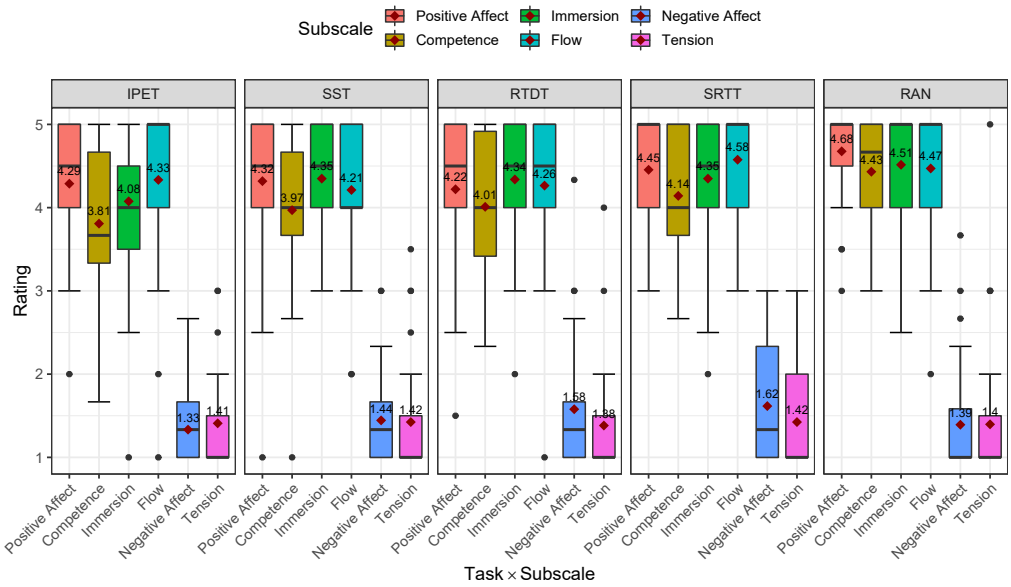


Figure 5.3: Children's game experience ratings by screening task (facet) and subscale (color). Diamond shapes represent mean values.

ure 5.2, the user experience was positive in all subscales for all screening tasks, which indicates that the classified benchmark categories were positive for all subscales (at least above average) and that all mean values were significantly above 0.8, representing a positive evaluation (cf. Schrepp, 2018).

Attractiveness of each screening task was rated as excellent. Regarding *Ease of Use*, children rated the RTDT and RAN as excellent, the IPET and SST as good and only the SRTT's ease of use was perceived as above average. As for the *Perspicuity*, the RAN received excellent ratings, the SST good ratings and the IPET, RTDT, and SRTT above average ratings. Lastly, children reported excellent *Stimulation* in the SST and RAN tasks, good stimulation in the RTDT and SRTT task and above average in the IPET task.

As for the differences in perceived user experience with respect to grade and literacy skills, we only found that second graders rated the RAN task as significantly less easy to use ($M = 1.55,\ SD = 1.29$) than third graders ($M = 2.48,\ SD = 1.16$), $p = .028$. We did not find significant differences for the other subscales and tasks between grades and literacy skills, p's $> .05$.

5.4.2 Game experience

To evaluate the children's game experience, we used a conservative approach to analyzing each subscale by conducting one-sample Wilcoxon signed-rank tests against the middle value of the subscale's 5-point Likert scale (3 = moderately). Descriptive results are displayed in Figure 5.3.

Children's ratings of each screening task on the subscales *Positive Affect*, *Competence*, *Immersion*, and *Flow* were significantly higher than moderately, p's $< .001$. In contrast, ratings of the *Tension/Annoyance* and *Negative Affect* subscales were significantly lower than moderately, p's $< .004$.

We found significant differences in children's ratings of the subscales *Positive Affect*, *Competence*, *Immersion*, for the RTDT between children with poor and typically developed literacy skills, p's $< .04$. Interestingly, children with poor reading and spelling skills perceived a significantly higher *Positive Affect* and reported feeling significantly more competent and immersed in the RTDT task compared to typically developing children, indicated by significant negative correlations between reading or spelling skills and these subscales, see Figure 5.4.

Further, third graders perceived significantly higher *Competence* in the IPET ($M = 4.06$, $SD = 0.78$) compared to second graders ($M = 3.23$, $SD = 0.82$), $p = .016$. Besides these differences, we did not find further significant effects of grade and literacy skills on other subscales and tasks, p's $> .05$.

5.4.3 Test situation

To investigate if children perceived the screening tasks more as games or as exams/tests, we evaluate the items of the TSQ descriptively and analyze each item by conducting one-sample Wilcoxon signed-rank tests against the middle value of the scale (0 = "neither game/test nor no game/no test" for the items using the bipolar rating scale and 3 = moderately for the items using the 5-point Likert scale). Descriptive results are displayed in Figure 5.5 and 5.6.

The results on whether the children perceived the individual screening tasks as a game/no game or a test/no test are displayed in Figure 5.5. The average rating of whether the individual screening tasks were perceived as a game or no game was significantly above 0, p's $< .001$, indicating that all screening tasks were more likely considered a game than

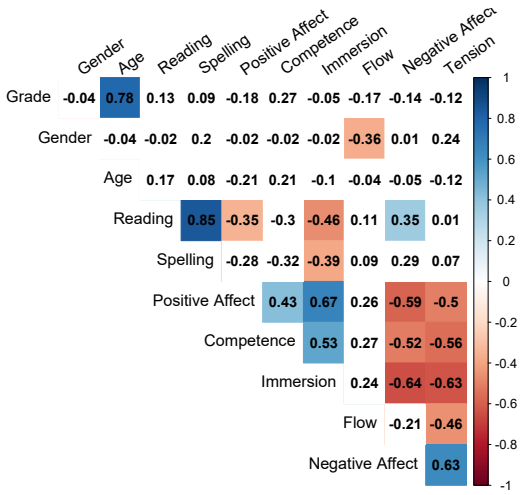


Figure 5.4: Spearman's Rank correlation for the RTDT between Kids-GEQ subscales, grade, age, and literacy skills. Correlations significant on $\alpha = .05$ are colored. One child was excluded from this analysis because their reading score was not included in the score table.

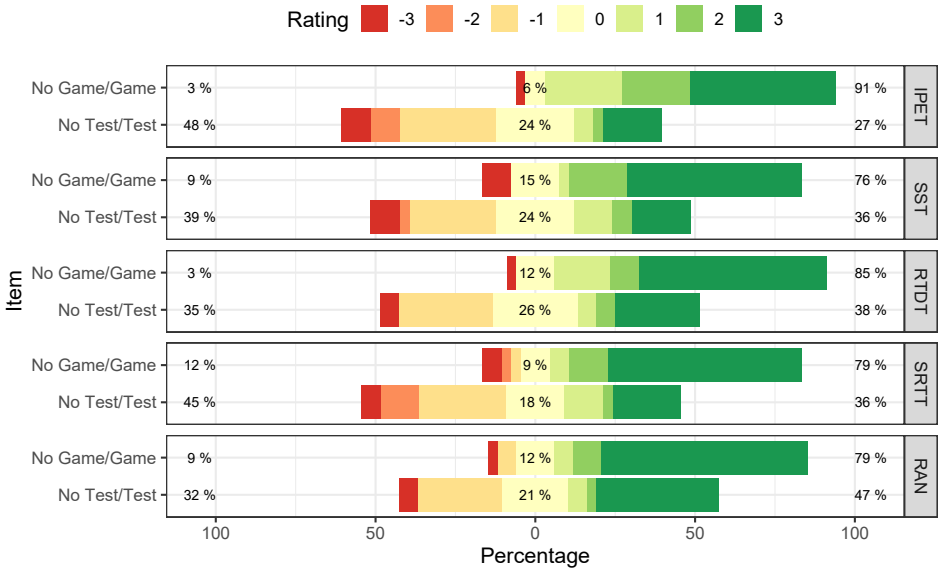


Figure 5.5: Children's ratings of the screening tasks (facets) for the items *No Game/Game* and *No Test/Test*.

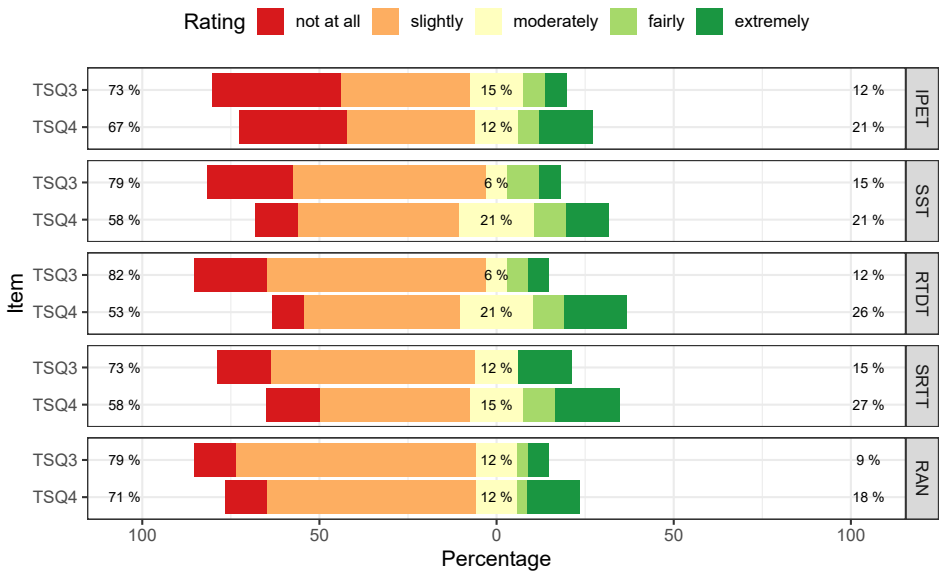


Figure 5.6: Children's test situation ratings of the screening tasks (facets) for the questions *"I was afraid during the exercise that I would do something wrong"* (TSQ3) and *"I felt like I was taking an exam during the exercise"* (TSQ4).

no game. The majority of the children perceived all screening tasks more like a game compared to no game, with at least 76% of the children tending to perceive the task as a game. In contrast, the test/no test item's average ratings were not significantly different from "neither test nor no test" for any screening task, p's > 0.226. While the IPET and SRTT were perceived as no test by more children than a test (48% vs. 28% and 45% vs. 36%, respectively), the RAN task tended to be perceived by more children as a test (47%) than no test (32%).

We did not find significant differences between grade and literacy skills on the ratings of these two items of the TSQ for any screening task, p's $> .05$.

The children's test situation perception results are displayed in Figure 5.6. The average ratings of whether children were afraid during the tasks do to something wrong (*TSQ3*) were significantly below moderately for all screening tasks, p's$< .003$. The majority of the children with at least 73% tended to report that they were not afraid of doing something wrong in the screening tasks. Children's ratings on whether they felt like taking an exam during the screening tasks (*TSQ4*) were significantly below moderate for the IPET, SST, and RAN, p's < 0.027, while the rating did not significantly differ

from moderately for the RTDT and SRTT, p's $> .052$. For all screening tasks, most children with at least 53% tended to report that they did not feel like taking an exam when they were working on the screening tasks, see Figure 5.6. Concerning the effect of grade and literacy skills on children's test situation perception, we found no significant differences, p's$> .059$. We only found significant differences between boys and girls in the RTDT for *TSQ4*, $p = .034$, and in the RAN for *TSQ3*, $p = .002$. That is, girls reported the RTDT felt less like an exam ($M = 1.88,\ SD = 1.68$) than boys did ($M = 3.06,\ SD = 1.68$) and boys reported in the RAN to be less afraid of doing something wrong ($M = 1.12,\ SD = 0.49$) than girls ($M = 2.24,\ SD = 1.39$).

5.4.4 Discussion

The results indicate an overall positive perception of the screening tasks (*Q2*) and prove their feasibility for use with primary school children (*Q1.1*) as well as the feasibility of the screening in groups (*Q1.2*).

The visual design and general implementation of the screening tasks appealed to the children, indicated by excellent ratings of their *Attractiveness* as well as high ratings of *Positive Affect*, *Stimulation*, and *Immersion*. Further, children reported perceiving high levels of *Flow* in each screening task, implying that they focused on the screening tasks and were not distracted by the game elements or the group testing situation. Additionally, children reported feeling competent while working on the tasks. Importantly, children did not perceive *Negative Affect* nor did they feel tense or annoyed during the screening tasks.

The high ratings of *Ease of Use* and *Perspicuity* indicate that the screening tasks are easy to use by the target group and prove the usability of the screening in that it can be used independently by the children with minimal adult instructions. The interactive tutorials and the one-way-only navigation approach very likely played a crucial part in this positive evaluation.

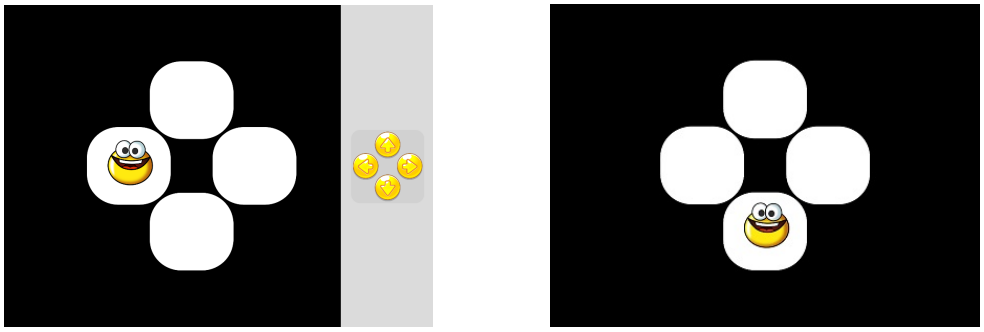
The screening tasks were also categorized as games rather than no games, while the children did not clearly classify the tasks as tests or no tests. The children might be primed beforehand by the situation and anticipation for the upcoming tasks, leading to uncertainties in answering the latter question. Importantly, children were not afraid of doing something wrong during the screening and did not feel like they were taking an exam. At the same time, they were able to complete the screening tasks, which may

also imply that the playful approach may ease potential test anxiety as seen in other game-based assessment studies (e.g., Kiili and Ketamo, 2018).

We found very few significant differences in children's perception of the screening tasks with respect to grade and literacy skills. We only found significant differences between children with poor literacy skills and typically developing children in the RTDT for which – in contrast to what one may expect – children with poor literacy skills perceived significantly higher *Positive Affect*, *Competence*, and *Immersion* than typically developing children. This further indicates that the overall look and feel of the screening and the implementation of the individual screening tasks are independent of children's literacy skills. These results might be explained by the fact that children with poor literacy skills usually perform worse on these tasks than their classmates and thus experience them negatively. In the current study, the game-based design may have not only mitigated but also motivated the potentially higher negative emotional experiences of children with poor literacy skills. Typically developed children may not perceive such tasks as negative at all, as they are expected to perform relatively well anyway. Thus, they are not as receptive to the game-based design.

Further, we only found a significant effect of grade in the *Ease of Use* in the RAN task as well as in the perceived *Competence* in the IPET tasks, for which second graders reported lower ratings than third graders

In addition to the overall positive children's experiences with the screening and as we already mentioned in the section about the SRTT construction, we observed that the thumb-pad input implemented for this task was not optimal. We observed that children had to pay a lot of attention to hit the correct arrow with their thumbs and were constantly looking at their thumbs rather than paying attention to the locations of the stimuli. This finding is supported by children's perceived *Ease of Use*, which was lower for the SRTT than for all other tasks but can still be considered above average. In contrast, the other tasks' *Ease of Use* scored good or excellent benchmark categories. This might be because the thumb-pad had no haptic feedback and the arrow buttons did not physically protrude, as for game controllers. In response, we changed the input of the SRTT to use direct touch input: instead of the thumb-pad at the border of the screen, the children directly touch the stimuli (yellow agent) in its location, resulting in a "catch" game, see Figure 5.7.



(a) SRTT version used in the present study that uses a thumb-pad with four arrow buttons as input.

(b) Revised SRTT version in which we have replaced the thumbpad with a "catch the agent" input where children click directly on the agent.

Figure 5.7: Two different designs of the SRTT. Figure 5.7a displays the version used in the second pilot (*P2*) study and Figure 5.7b the revised version. We applied changes as a response to the results and observations of the feasibility study, which highlighted children's problems with the non-haptic thumb pad.

5.5 Conclusion

In this chapter, we presented the design and feasibility evaluation of the present screening.

First, we elaborated on the design and content of the screening, underlining significant design choices that make the digital screening easily accessible and engaging for young children aged 5-10 years. These choices facilitate its application in group settings, answering the Research Question RQ1.1 for this thesis: "Can children use the screening independently, and how do they perceive it?". Our findings indicate that children can indeed utilize the screening independently, with minimal adult supervision required. The incorporation of interactive tutorials and intuitive navigation contributed significantly to this independence, ensuring that even pre-reader age children found the screening simple to understand and use. Moreover, the integration of game elements led to a positive reception of the screening, with children engaged and motivated to complete it, thus affirmatively addressing the concerns of RQ1.1 regarding user experience and engagement.

As we have previously argued, conducting screenings in group settings offers significant

advantages in terms of cost and time efficiency compared to individual assessments. This approach directly relates to RQ1.2: "Is the presented literacy screening applicable in a group setting?" Our evaluation in the second pilot (*P2*) with 34 German second and third graders aged 8-10 years has provided evidence that the screening is indeed feasible and effective in group settings. Optimized seating arrangements, the use of headsets, and the implementation of engaging elements allowed for the management of up to 10 children by a single test administrator. These adaptations ensured focused work in the group environment, affirmatively addressing RQ1.2 by demonstrating the screening's applicability and efficacy in group settings.

In addition, the positive perception of the screening tasks as games rather than tests was particularly beneficial for children with test anxiety, indicating a differentiated understanding of the way screening is received by different groups of children. This aspect not only contributed to a high usability rating but also highlighted the potential of such digital screenings to create a more inclusive and less intimidating evaluation environment. Notably, children with poor reading and spelling skills found the tasks equally or more positive than their typically developing peers, suggesting that the screening's design successfully mitigates negative feelings towards assessments.

In conclusion, the results from our pilot study affirmatively address both research questions RQ1.1 and RQ1.2. The screening's design and digital implementation enable independent use by children, offering a positive and engaging user experience that supports its feasibility in group settings.

Table 5.3: Lessons learned after conducting the second pilot study (*P2*) in designing group-based pre-reader literacy screenings.

Category	
Group testing	• Use a comfortable and child-friendly headset when audio in- or output is required • Use highly interactive tutorials for each task • When multiple tasks are used, implement (game) elements that (i) provide relaxation and (ii) keep children engaged to avoid distracting other children while they wait for other children to complete a task • Conduct multiple pilot tests at early stages to determine (i) technical constraints, (ii) children's social behavior in group settings
Design	• Avoid input that is not intuitive for the modality, such as a virtual thumb-pad for tablets

Chapter 6

Methods of the Screening Study

The R scripts and datasets supporting the findings of the screening study are openly available and hosted on the Open Science Framework[14].

6.1 Study design

The screening study was set up as an experimental field study in a longitudinal design within a natural setting. Data were collected at the schools twice. The first data collection point (T1) took place in the first grade shortly after school enrollment in the fall of 2018. The children performed the tasks on a tablet and collected the intelligence measurement on paper and pencil. The second data collection was conducted with the same children in the middle of the second grade in early 2020 (T2). In addition to the intelligence measurement, the SLRT II spelling test and the SLS 2-9 reading test were also collected with pen and paper. Finally, to build predictive models for reading and spelling performance, we took the collected reading and spelling data from T2 - unavailable in T1 as the children have just entered school - as ground truth data and had the models trained on the T1 data.

6.2 Recruitment and participants

For the recruitment, the principals of 60 schools from the districts of Tübingen, Reutlingen and Zollernalbkreis, Baden-Württemberg, were contacted after consulting the study

[14] `https://osf.io/hrbdv/?view_only=336b53066cfc4057b23043be750a2fb9`

with the regional school council and an ethics committee. Upon obtaining consent from 33 schools, the children's parents were informed about the aim of the study and the schedule and method of data collection. Parents gave their consent for their children to participate in the screening and for audio files to be recorded during the screening. Participation was voluntary and could be terminated at any time without giving reasons. As compensation, the parents received free access to the mobile training app "Prosodiya - Lesen und Schreiben lernen" for their children after completion of the study. On the day of testing the children were allowed to choose an eraser, pencil or sticker as a participation reward. In an effort to recruit as many children as possible, there were no eligibility requirements other than the fact that they were attending first grade at the time.

521 German primary school children from first grade registered for the screening study. 107 children eventually had to be excluded from the screening study due to one of the following reasons (also see Figure 6.1):

- missing registration data
- missing at least one test date (T1 or T2) due to sickness, relocation of home, change of school or absence of leave
- deregistration by a parent
- canceling or interrupting the screening procedure
- test problems and test misconduct, perceived by testers (for example, deliberate random clicking, didn't understand the instructions, visible lack of motivation)
- technical issues
- data was used for the final pilot test (P3)

The missing age of one child and missing birthday of another child could be imputed by calculating the median age across all other children, leaving a total of 414 children (205 girls) aged $6-9$ years ($70-111$ months, $M = 79.76$, $SD = 4.80$) at T1 and $7-10$ years ($85-126$ months, $M = 95.43$, $SD = 4.82$) at T2 for further analysis. At the time of registration, none of the children had a diagnosed reading and spelling disorder, and 29 children were in speech therapy, 11 in occupational therapy and 10 in learning therapy. Data from these children were included because of the generalizability of the screening. 84.06% of the children (N=348) were German native speaker. Figure B.1 shows the classification of children's literacy skills into groups of weak and non-weak. We describe the calculation for this grouping in Section 2.1.1. When this calculation is applied for

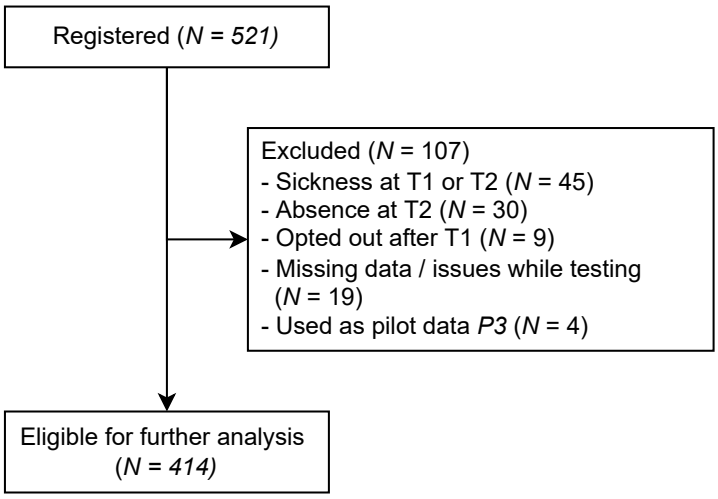


Figure 6.1: Composition of study participants and list of reasons for exclusion.

both literacy skills separately, it shows that 16.87% of children are weak readers, 22.98% are weak spellers, 29.10% are either weak readers or spellers and 10.76% are weak in both reading and spelling.

6.3 Materials

We used 30 Samsung Galaxy Tab A 2016 tablets (type SM-T580) with 10.1-inch screen size and a resolution of 1920×1200 pixels running Android 7.1 in the main screening study. We put the tablets on tablet stands from Ikea (type ISBERGET) that offer a 45° viewing angle.

We connected the tablets to NUBWO N2 headsets with suspension-style headbands that offer a flexible and comfortable hold. We fixed the integrated volume and microphone controls with adhesive tapes so that the children could not accidentally mute the speakers or microphones.

6.4 Screening procedure

In order to conduct the screening at schools, one instructor or teaching staff prepares the hardware, assigns seats, and gives general instructions about the screening procedure. When the children were seated at their desks, they were asked to put on the headphones for the duration of the screening and listen carefully to the instructions given by the

tablet. The children could then start the screening individually whenever they felt ready.

6.5 Measures

6.5.1 Nonverbal intelligence

Many intelligence tests still require reading and spelling skills, which pre-selects the people to be tested and excludes young children, illiterates and people with poor test language skills. Consequently, nonverbal intelligence tests were developed and are now widely in use. By making the tasks nonverbal and not using speech in any capacity, culturally appropriate tests aim to measure participants' fluid general intelligence and thus can be completed regardless of prior knowledge and status or educational influences (Gold, 2018). Within the context of this doctoral thesis, we have developed such a test for measuring the intelligence indicator of children. The standardized test we used as a basis is a six-part commercial test developed for children aged 5;4 to 9;11 years old, which takes about 45-60 minutes. With split-half reliability of $r = .90$ (N = 194), the subtest of this commercial test named 'Series Continuation' is considered a reliable intelligence index measure and takes only about 8-10 minutes. Based on this subtest and due to limited resources, processing time and test feasibility, we developed our own test to measure an intelligence index called Continuous Series Test (CST). The CST is structurally similar to the commercial, standardized test subtest: it consists of 15 tasks showing a series of three figures and requiring the identification of a target among four distractors, which logically continues the series. However, since the CST is a subtest of a complete intelligence test, only an indication of intelligence can be derived from the results. Reents (2020) was the first to demonstrate the convergent validity of the CST with $r = 0.76$.

In the instructional manual, the authors of the commercial test suggest that the results be evaluated in two blocks of three subtests each. Since the CST is based only on one subtest of the test, we calculated the test score for the CST by summing the correct responses.

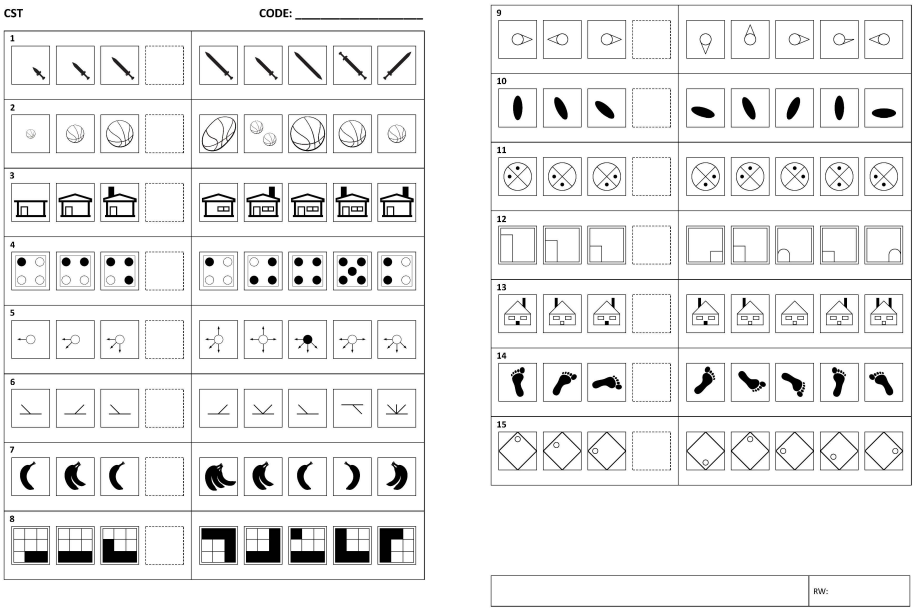


Figure 6.2: The CST is a two-sided, nonverbal, culturally appropriate test designed and administered to measure children's intelligence index.

6.5.2 Reading and spelling

Measurement of reading skills. The Salzburger Lese-Screening for Grades 2-9 (Mayringer and Wimmer, 2014) (SLS 2-9) was used to examine reading ability. The SLS 2-9 is an economical and reliable single-test procedure for identifying students with weaknesses in basal reading skills. Basal reading weaknesses are primarily indicated by significantly reduced reading speed. The SLS 2-9 consists of 100 sentences whose statements are either true or false, i.e., 'Snow is red'. The children had 3 minutes to read the sentences and evaluate by circling a tick for a true or circling a cross for a false statement. The number of correctly judged sentences was evaluated and used as a raw score in the data analysis. We used the parallel versions A1 and A2 alternately for the respective seat neighbors as an anti-cheat measure. The forms differ only in the order of the sentences. The SLS 2-9 is a paper-pencil test. According to the authors of the test, parallel test reliability is $r = .95$ (N = 107) for the second-grade level. We computed the SLS 2-9 raw test score by deducting the number of errors and omitted items from the total number of solved items. As a result, the minimum SLS 2-9 score is 0, while the maximum SLS score is 100.

Measurement of spelling skills. The Salzburger Lese- und Rechtschreibtest 2nd revised edition (Moll and Landerl, 2010) (further SLRT II) is a German reading and spelling test for children in grades 1-6 that detects issues in written language acquisition. It consists of both a reading fluency and a spelling subtest. However, for the present study, we only used the spelling subtest. The spelling test assesses the competence to insert dictated words orthographically correctly into frame sentences. The children are supposed to complete 24 gap sentences, each sentence missing one word. The experimenter reads the target word out loud three times: once before, once as part of and once after the entire sentence. We used version A of the SLRT II in paper-pencil testing. According to the authors, the retest reliability for the spelling subtest for the total number of errors is $r = 0.85$ (N = 40, retest interval five weeks) for the second school level and version A. Overall, the SLRT II takes 20 to 30 minutes. The SLRT aims to identify intervention needs and does not sufficiently distinguish between average and above-average performance. In a deviation from the manual's suggested scoring of different types of errors, we calculated the SLRT II raw test score by subtracting the number of errors from the total number of items solved. Therefore, the minimum SLRT II raw score is 0 and the maximum SLRT II raw score is 24.

6.6 Task-specific analysis

For each task while testing, a comma-separated-values log file was created locally on the tablet and stored in a separate folder with the subject ID as the folder name. Each line in one of these log files represents a task trial that records information depending on the task. Table A.4 lists all of this information.

Considering these parameters, we report the planned statistical analyses of each task in the following subsections. The goal was to assess the tasks' validity and define predictors for further machine learning analysis.

6.6.1 Incidental Holistic Perception Task (IPET)

As part of the screening project, Moschko (2018) analyzed the Incidental Holistic Perception Task (IPET) with a small sample ($N = 34$) from the P2 pilot (see Table 4.1). We split our analysis of the task into two parts. In the first part, we applied the same methods as Moschko (2018) used in the IPET pilot evaluation. This approach is briefly

summarized hereafter. In the second part, we further investigated children's task performance using measures from signal detection theory.

We investigated whether there are systematic differences with respect to certain object properties, such as object type (fictitious vs. real) and object class (distractor vs. target object). The IPET performance results were analyzed using a 2x2x2 factorial experimental design, where the group (weak reader/writer and non-weak reader/writer) represents a two-level between-subject factor, and object type and object class each a two-level within-subject factor. We used a generalized linear mixed model from the R package *lme4* (Bates et al., 2015) to analyze the influence of the factors group, object type and object class on recognition performance. As a group, we defined the children with and without literacy difficulties according to our criteria defined in Section 2.1.1, being one standard deviation below the mean of the reading or spelling raw score.

We used additional metrics from signal detection theory to gain further insight into children's performance, especially considering the different literacy skill categories. Besides d' as a performance measure, we calculated the response bias c. We used those measurements for the overall task performance and the specific object type (real or fictitious stimuli). By using d', we improve the informative value of task performance by taking into account response tendencies, e.g., children who always clicked the same side. A high, positive d' value indicates good performance, a value of zero means that the child answered randomly, whereas a negative d' may indicate that the child did not understand the task or was confused. In order to obtain d' and c, we calculated the hit rate (*hr*, correctly identified targets), miss rate (*mr*, rejected targets), correct rejection rate (*crr*, correctly identified distractors) and false alarm rate (*far*, rejected distractors) for each individual and all answers given, as well as for both object types. We did this for weak readers and spellers separately and for children who are weak in both skills. To correct for a bias in extreme values in the hit and false alarm rate, we have applied the log-linear correction approach according to Hautus (1995):

$$\begin{aligned} hr &= \frac{\Sigma hits_{target} + 0.5}{n+1} \\ crr &= \frac{\Sigma hits_{distractor} + 0.5}{n+1} \\ mr &= 1 - \frac{\Sigma hits_{target} + 0.5}{n+1} \\ far &= 1 - \frac{\Sigma hits_{distractor} + 0.5}{n+1} \end{aligned} \tag{6.1}$$

Given that information, we calculated the sensitivity measure d' and response bias measure c, also known as criterion location, using the z-transformations of given values (Macmillan and Creelman, 1990):

$$\begin{aligned} d' &= z(hr) - z(far) \\ c &= -0.5(z(far) - z(hr)) \end{aligned} \tag{6.2}$$

After we excluded 107 children from the screening in general due to missing and incomplete data as described in Section 6.2, we excluded another 14 children specifically for the IPET analysis. For a more detailed list of the reasons for exclusion, see Table B.1. The final sample of the task analysis consisted of $N = 400$ children.

6.6.2 Syllable Stress Task (SST)

The basis for the evaluation of the Syllable Stress Task (SST) was laid by Neubrand (2020) in her bachelor thesis. As part of the screening project, she had a major portion of the data ($N = 354$) at her disposal for her analysis. However, because there were differences in the exclusion of children and thus in the underlying data set, we repeated their analysis to see if similar results could be found. Using multiple linear regression,

Table 6.1: For each target variable (tv) we created and compared these models, including sum score, sum of clicks and their interaction as possible predictors. If covariates (cv) were part of the model, we included them before the predictors as interaction terms (age * gender * CST performance). The models one to three are part of *model set 1*, whereas for *model set 2*, the sum score is replaced with the person parameter.

Model 1:	tv $\sim$ (cv) + sum score (ss)
Model 2:	tv $\sim$ (cv) + sum score (ss) + sum of clicks (soc)
Model 3:	tv $\sim$ (cv) + sum score (ss) * sum of clicks (soc)

we created classical test theory models with sum score and total button clicks to repeat the audio of target stimuli (sum of clicks) as predictors (*model set 1*, see Table 6.1). We further compared which model made the best predictions for children's reading skills and which made the best predictions for their spelling skills using likelihood ratio tests. In addition, we created and compared four-item response theory (IRT) models. These models were set up with the *mirt* package (Chalmers, 2012). Starting from the Rasch

model, each of these contained one more parameter to be estimated than the previous model, resulting in the sequential addition of a Rasch-model with guessing, 3-parametric logistic model (3PL) and a 4-parametric logistic model (4PL). In Table B.6, the item characteristic function is given for each model. We generated the individual person parameters from the model that best described the data, also found by likelihood ratio tests. We then tested the predictive performance of models (*model set 2*) with the person parameters and the total button clicks to repeat the output of target stimuli as a predictor of literacy skills of the children, similar to the analysis of models from *model set 1*. We conducted model tests with and without covariates for both classical test theory and person parameter based approaches, as shown in Table B.6. We also checked models without covariates to investigate the sole task performance of predictors and see their influence on literacy predictions.

Finally, we compared the best models from both model sets using the Bayesian Information Criteria (BIC) score and r^2_{adj} to see which ones work best as predictors for each literacy skill.

After we excluded 107 children from the screening in general due to missing and incomplete data as described in Section 6.2, we excluded another 11 children specifically for the SST analysis. For a more detailed list of the reasons for exclusion, see Table B.2. The final sample of the task analysis consisted of $N = 403$ children.

6.6.3 Rising Time Discrimination Task (RTDT)

Similar to the evaluation of the syllable stress task, Wölfl (2020) used multiple regression models in her bachelor's thesis to examine the impact of children's performance in the Rising Time Discrimination Task (RTDT) on their literacy skills. A substantial amount of screening data ($N = 287$) was available to the author, whereas given the variation in exclusions of children and differences in the underlying data set, we performed a repeated analysis to determine whether comparable results could be obtained. We built multiple linear regression models based on classical test theory, using children's sum score and their total amount of clicks on the repeat instructions button (sum of clicks) as predictors (*model set 1*, see Table B.6). We tested their predictive performance concerning children's literacy skills. Analysis was done with and without covariates to test task performance's sole predictive power. We compared the models using likelihood ratio tests to find the best-performing model. Furthermore, we developed and compared

various Rasch models based on latent trait theory using likelihood ratio tests. We next constructed the individual person parameters based on the Rasch model that best described the data. We tested the prediction performance of models with the person parameters and the total amount of clicks on the repeat instructions button as predictors (*model set 2*) to predict the children's literacy skills. This approach is therefore similar to the tests done with the classical test theory models. Finally, we compared the best models from both models sets using the Bayesian Information Criteria (BIC) score and r^2_{adj} to determine which works best as predictors for each literacy skill.

After we excluded 107 children from the screening in general due to missing and incomplete data as described in Section 6.2, we excluded another 43 children specifically for the RTDT analysis. A large portion of those children answered the items systematically during testing. Here, we considered whether the children always chose one side or systematically alternated left-right or right-left. For a more detailed list of the reasons for exclusion, see Table B.3. The final sample of the task analysis consisted of $N = 361$ children.

6.6.4 Serial Reaction Time Task (SRTT)

As part of the screening analyses, Metelmann (2020) was able to evaluate the Serial Reaction Time Task (SRTT) with $N = 287$ of the screening participants. She used studies by Lum et al. (2010) and van der Kleij et al. (2019) as a basis for her evaluation of the data. Similar to the previous task analysis, we expect the underlying data set to be different since the exclusion rules of children varied. We therefore repeated the analysis by Metelmann (2020).

The SRTT evaluation with grouping (based on Lum et al. (2010) and van der Kleij et al. (2019)) is a 5 (block) x 2 (group) mixed-factor design comparing blocks within subjects and groups. The evaluation was split into the analysis of the accuracy and the reaction time (see Section 4.4.4 for detailed reasoning). For both parameters, we have implemented the respective analysis from Lum et al. (2010) and van der Kleij et al. (2019) with slight adjustments.

Table 6.2: Planned evaluation to calculate predictors from accuracy and reaction times for the SRTT

Based on Source	Statistical Analysis
Accuracy	
Lum et al. (2010)	We fitted a generalized linear mixed effects model with the blocks, group and their interaction as fixed effects, as well as a random intercept. As target variables, we used a matrix of correct and incorrect trials per child. To avoid arc transformation and to account for the binomial distribution of error frequencies, we used this analysis instead of mixed anova, as the authors did. Differences between individual blocks were considered via Bonferroni-corrected pairwise post-hoc tests using the *emmeans*[15] package.
van der Kleij et al. (2019)	We compared the groups in terms of error rates using a t-test for independent samples. To do this, error rates were averaged first for each child per block and then per child across blocks.
Metelmann (2020)	Complementing the Lum et al. (2010) and van der Kleij et al. (2019) evaluation, the correlation was calculated to investigate whether there was a trade-off between reaction time and error rates among children.
Reaction time	
Lum et al. (2010)	We calculated an analysis of variance with the reaction time difference between blocks four and five as the dependent variable, and the group as the independent variable. Further, we set up a variance analysis model that included the reaction times of all blocks as the target variable and the group, the blocks, and their interaction as independent variables.

[15]Estimated marginal means (emmeans) package for r: `https://github.com/rvlenth/emmeans`

Table 6.2: Planned evaluation to calculate predictors from accuracy and reaction times for the SRTT

Based on Source	Statistical Analysis
van der Kleij et al. (2019) & Metelmann (2020)	We used piecewise growth models with a comprehensive random effects structure to analyze reaction times, following van der Kleij et al. (2019). Based on the suggestions by Metelmann (2020), we adjusted the coding for learning curves (0 to 4) and increased the response (coded as 0 until the fourth block, then 1) to refine comparisons. Linear mixed-effects models were fitted for each child, incorporating response increase, learning curve (linear, squared, cubic), and their interactions, along with an intercept and a random slope. Model comparison was done using variance analysis, in line with van der Kleij et al. (2019).
Further analysis	
Metelmann (2020)	Our groups were based on reading and spelling tests and were not independent, thus the group variable was excluded from our literacy prediction models, following the approach of van der Kleij et al. (2019). We used piecewise growth models without the group variable, selecting the best model based on BIC, as recommended by Metelmann (2020). The coefficients from the best reaction time model, along with accuracy, were used to identify the most predictive models for reading and spelling. We began with complex models incorporating all features, then iteratively simplified them by removing the least significant predictors based on their p values, continuing until only significant predictors remained. This procedure was applied to models both with and without covariates.

We have divided the analysis into two scenarios, since a different data basis emerges depending on the exclusion criteria.

Scenario 1: Exclusion similar to previous studies Exclusions of children were based on a total of more than 50% of the trials being incorrectly clicked in the SRTT. The exclusion rule is based on suggestions by Hsu and Bishop (2014) in including children for further analyses only if they scored at least 70% correct, to ensure that the task was properly understood and seriously addressed. We adjusted the threshold from 70% to 50% for several reasons. Since our test setup consisted of four stimuli instead of three, the margin of error was significantly higher, especially for children with reading and writing difficulties. Furthermore we expected the task to be more intuitive to use, because children click directly on the stimulus and no additional medium is needed. Contrary to van der Kleij et al. (2019), we applied the 50% criterion for both the reaction time analysis and the accuracy analysis, as we saw no reason not use the same data basis. In total, this rule lead to the exclusion of 18 children.

Scenario 2: No exclusion of accuracy performance outliers To find the best predictors, we wanted to keep the data basis as broad and general as possible. Since individual outliers were unlikely to significantly bias the data, we made no exclusions for accuracy and, consequently, reaction time performance outliers.

After we excluded 107 children from the screening in general due to missing and incomplete data as described in Section 6.2, we excluded another 31 children specifically for the RTDT analysis. An additional 18 children were excluded for the scenario 1 analysis, as mentioned earlier. For a more detailed list of the reasons for exclusion, see Table B.4. The final sample of the task analysis consisted of $N = 365$ children.

6.6.5 Rapid Automatized Naming Task (RAN)

Since speech recognition and automatic editing of audio files are very time-consuming with a high probability of producing inaccurate and incomplete results, we manually edited the RAN data for each child. During the manual evaluation, we had to correctly mark the beginning and end of the audio recording to get the total length of the audio file and to note if the child omitted words or misnamed the given image. We marked those mistakes and used the generated information to calculate the following possible predictors: time per item, amount of type 1 errors (wrong-naming) and amount of type 2 errors (omitted-words).

In the evaluation, first we fitted a linear model with page as a predictor for itemtime to investigate sequence effects between pages. Secondly, we examined itemtime differences between groups across all pages using a t-test, to check for group differences in overall performance. We further set up several linear regression models for predicting reading and spelling skills that included itemtime, wrong-naming error and omitted-words error as predictors together with covariates age and CST performance. We setup the models, starting with the most complex model containing all of these predictors. The models were reduced stepwise, removing the predictor with the lowest contribution to prediction at each iteration, characterized by the p value. This process is repeated until only significant predictors are left in the model. This analysis is applied to model variants, with and without covariates.

After we excluded 107 children from the screening in general due to missing and incomplete data as described in Section 6.2, we excluded another 68 children specifically for the RAN analysis. A large portion of excluded children had defective audio recordings or completed only one of the two pages. However, we excluded no children or trials of children due to outlying reaction time behavior. As research shows (Ulrich and Miller, 1994), such truncation may introduce more unwanted bias. For a more detailed list of the reasons for exclusion, see Table B.5. The final sample of the task analysis consisted of $N = 346$ children.

Chapter 7

Task-specific Results

In this section, we look at the results of the screening tasks evaluation. Our analysis aims to provide insight into these tasks' ability to predict literacy skill development and determine predictors, that help to do so. We also look at the importance of covariates in those analyses.

7.1 Incidental Holistic Perception Task (IPET)

Each child completed 51 trials in total, with three trials excluded as they were training trials. We analyzed the two image sets for independence, one of which was always shown to the children, depending on their seated position in the test session. There was no significant effect for better recognition of images from set one compared to set two, $t(19167) = 1.27$, $p = .205$, despite the average accuracy being slightly greater in set one ($M = 0.5180$, $SD = 0.4997$) compared to set two ($M = 0.5088$, $SD = 0.4999$).

Starting from the basic intercept model, we added factors step by step and examined how the prediction of recognition performance changes across all other conditions. The order in which factors were added can be seen in Table 7.1. Adding group as a factor did not seem to improve the prediction ($\chi^2(1) = 0.04$, $p = .842$, $BIC = 26641$). The same was true for adding the object type ($\chi^2(1) = 0.48$, $p = .489$, $BIC = 26651$), the interaction between group and object type ($\chi^2(1) = 0.29$, $p = .592$, $BIC = 25947$) and the interaction between group, object types, and object class ($\chi^2(1) = 0.15$, $p = .698$, $BIC = 25756$). Adding object class significantly improved prediction ($\chi^2(1) = 723.00$, $p < .001$, $BIC = 25937$) as well as adding the interaction of group and object class

Table 7.1: Parameter estimators for the considered predictors of the model that best describes the data, along with the 95% confidence intervals.

	Estimator	95% CI	p
Intercept	-0.11	[-0.17, -0.038]	**.002**
Group	-0.28	[-0.40, -0.16]	**<.001**
Object type	-0.33	[-0.42, -0.23]	**<.001**
Object class	0.29	[0.21, 0.38]	**<.001**
Group:object type	-0.04	[-0.17, 0.10]	.622
Group:object class	0.59	[0.45, 0.73]	**<.001**
Object type:object class	0.73	[0.61, 0.84]	**<.001**

($\chi^2(1) = 68.97$, $p < .001$, $BIC = 25888$) and the interaction of object type and object class ($\chi^2(1) = 151.38$, $p < .001$, $BIC = 25746$).

The model parameters of the last significant model, which included the interaction of object type and object class, proved best for describing the data, with BIC being the lowest of all compared models at $BIC = 25746$. The estimated model parameters with the corresponding confidence intervals for this model are shown in Table 7.1. All estimate values had significant influence, except the group and its interaction with object type.

Signal detection theory measures are displayed and summarized into literacy categories and object types in Table 7.2. Since we calculated miss rate (mr) and false alarm rate (far) by $1 - hitrate$ (hr) and $1 - correctrejectionrate$ (crr) respectively and their informative value with regard to the performance did not differ thereby, we excluded them in this summary. We also included the average accuracy performance for comparison for each literacy group.

7.1.1 IPET discussion and summary

Contrary to expectations, but in line with Moschko (2018) findings, the IPET did not reveal a significant difference in performance between the groups. Overall, both weak and non-weak literacy children recognized approximately the same number of objects correctly, as seen in the overall performance measure d' of all groups. Accuracy group

Table 7.2: Summary of signal detection theory measures, further split into object types and literacy categories. *Lrs.cat* 0 refers to children we considered to have no reading and spelling difficulties, 1 to children with only reading difficulties, 2 to children with only spelling difficulties and 3 to children we considered having both. For children from category 0, the response bias c is constantly lower compared to the other groups.

							real				fictitious			
lrs.cat	n	acc	hr	crr	d'	c	hr	crr	d'	c	hr	crr	d'	c
0	306	0.51	0.59	0.44	0.10	-0.30	0.63	0.39	0.12	-0.40	0.54	0.47	0.06	-0.13
1	28	0.52	0.67	0.36	0.14	-0.54	0.70	0.30	0.11	-0.64	0.62	0.41	0.15	-0.35
2	32	0.51	0.68	0.34	0.08	-0.62	0.70	0.30	0.07	-0.65	0.66	0.36	0.07	-0.48
3	34	0.51	0.62	0.40	0.12	-0.42	0.66	0.37	0.16	-0.51	0.56	0.44	0.09	-0.22

differences in identifying fictitious stimuli could not be found in contrast to the pilot analysis (Moschko, 2018), as Figure 7.1 shows the same course for both groups in both conditions.

The children's response strategies for fictitious items tended to be a mix of *yes* and *no* answers, whereas there was a stronger *yes* tendency for real objects. If we look at the group categories with literacy difficulties and their response bias c for each object type, we see a generally stronger *yes* tendency, indicated by the larger negative c value. This indicator seemed to somewhat differentiate between non-weak versus weak literacy children. In addition, we decided to include d' of each object type as a measure for performance, which seemed to be a more comprehensive measure than accuracy.

The measures from signal detection theory allow us to look at response bias and its interaction with task performance. However, the performance level of the children shows us that we must assume a high degree of guessing and that we cannot test the effects investigated initially. In this exploratory task, we see strategic differences (c) rather than differences in perception. To counteract guessing, at least the presentation time of the picture grids should be increased.

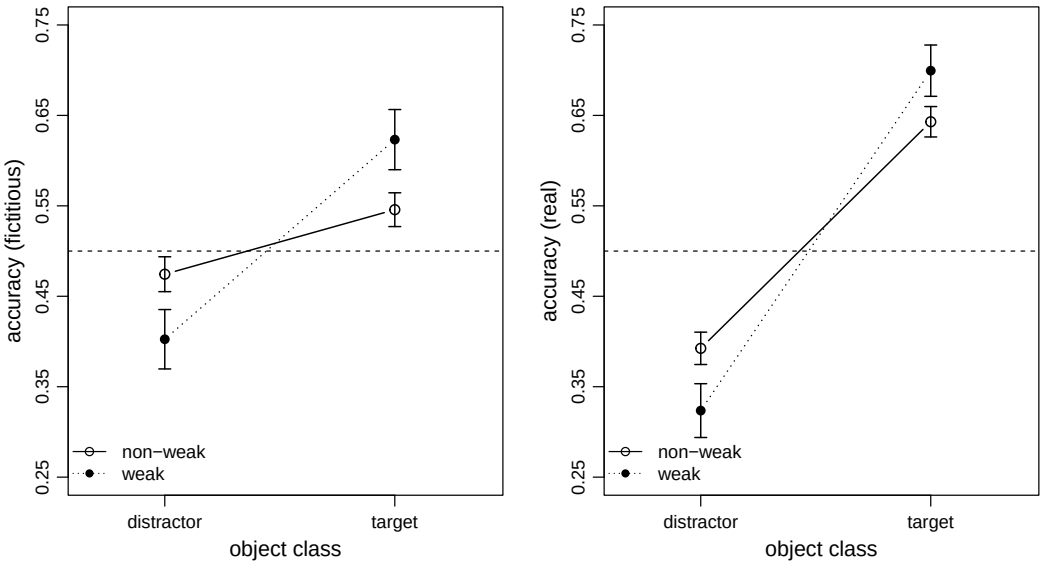


Figure 7.1: Accuracy of the respective groups for real and fictitious items when presented as targets or distractors. The weak and non-weak literacy groups each show a similar progression in recognition performance, contrary to previous findings by Moschko (2018).

Table 7.3: Overview of predictors resulting from IPET analyses adopted for further machine learning analyses.

Literacy skill	Predictors
Reading	- response bias c for real stimuli
	- response bias c for fictitious stimuli
	- performance indicator d' for real stimuli
	- performance indicator d' for fictitious stimuli
Spelling	- response bias c for real stimuli
	- response bias c for fictitious stimuli
	- performance indicator d' for real stimuli
	- performance indicator d' for fictitious stimuli

7.2 Syllable Stress Task (SST)

The following IRT models were set up and compared sequentially: Rasch model, Rasch model with guessing, 3PL model, 4PL model. A detailed description of these models can be seen in Table B.6. Results, displayed in Table 7.4, show the Rasch model with guessing as the best and most economical IRT model. From that model, we calculated the person parameter (pp), set up literacy prediction models as described in Table 6.1 and compared them in terms of prediction power. We found the correlation between the estimated person parameter and the sum score to be strongly correlated, $r(401) = 0.94,\ p < .001$. We also set up sum statistic models and tested which is best suited for prediction. The results for both literacy skills can be seen in Table 7.5. We performed ANOVA tests to compare models with the sum score (ss), sum of clicks (soc) and their interaction for predicting reading and spelling skills. We did the same with the person parameter (pp) that substitutes for the sum score. For both literacy skills, the simplest models, which only use the sum score or the person parameter, are also the most predictive. This finding is persistent with and without covariates taken into account. However, models including the covariates have a much higher amount of explained variance (r^2_{adj}). When we compare the models with only the sum score as predictors to the models with only the person parameter, we see that their BIC is close to equal. Assuming the rule of thumb for the BIC, as suggested by Burnham and Anderson (2004), the models are practically equivalent for both literacy skills, with their differences being smaller than two units. Taking a closer look at the person parameter model, we measured a significant contribution of the CST performance, its interaction with age and the person parameter on reading measure. Only the person parameter contributed significantly to predicting spelling skills. Estimators and confidence intervals can be seen in Table 7.6. These results don't change or improve when we look at parameter estimators for the sum score models.

7.2.1 SST discussion and summary

The sum score and the person parameter in the SST contributed to the children's later reading and spelling skills. These results confirm the findings of Sauter et al. (2012). Furthermore, the covariate CST and its interaction with age also contributed to children's reading ability. It made no difference how many clicks the children took to listen to the task again. The best-performing models' explained variance fraction is moderately

Table 7.4: Comparison of IRT models using ANOVA test statistic and the information criteria AIC and BIC. The Rasch model with guessing showed the best and most fitting results with a significant test statistic and the lowest AIC/BIC score.

	χ^2	p	AIC/BIC
Rasch	-	-	5754/5806
Rasch with guessing	$\chi^2(12) = 113.62$	**<.001**	5664/5764
3PL	$\chi^2(11) = 9.46$	.580	5677/5821
4PL	$\chi^2(23) = 20.02$	.641	5690/5882

Table 7.5: Results of SST data analysis show that the simplest models, which also include covariates age, gender and intelligence, perform better. The comparison of the two top-performing models using the BIC score shows virtually no difference in prediction power. ss=sum score, soc=sum of clicks, pp=person parameter.

		Models	Reading skill			Spelling skill		
			F-score	p	r^2_{adj}	F-score	p	r^2_{adj}
With	1.1	ss	-		10.08%	-		11.27%
covariates	1.2	ss+soc	$F(1) = 0.24$	.625	10.62%	$F(1) = 0.08$	.784	11.06%
	1.3	ss:soc	$F(1) = 0.79$	.376	10.58%	$F(1) = 0.41$	.524	10.92%
	2.1	pp	-		11.33%	-		11.20%
	2.2	pp+soc	$F(1) = 0.22$	.639	11.15%	$F(1) = 0.05$	.827	10.98%
	2.3	pp:soc	$F(1) = 0.67$	.415	11.07%	$F(1) = 0.08$	.775	10.76%
Without	3.1	ss	-		3.84%	-		7.40%
covariates	3.2	ss+soc	$F(1) = 0.02$	.891	3.61%	$F(1) = 0.03$	.872	7.17%
	3.3	ss:soc	$F(1) = 1.07$	.303	3.62%	$F(1) = 0.77$	.382	7.12%
	4.1	pp	-		4.27%	-		7.12%
	4.2	pp+soc	$F(1) = 0.01$	.934	4.04%	$F(1) = 0.00$	.948	6.89%
	4.3	pp:soc	$F(1) = 0.99$	.321	4.03%	$F(1) = 0.21$	.648	6.70%
Model	1.1	ss	$BIC = 2856$		10.08%	$BIC = 2419$		11.27%
comparison	2.1	pp	$BIC = 2854$		11.33%	$BIC = 2419$		11.20%

Table 7.6: Parameter estimators for the model with person parameter predicting spelling and reading skills, along with the 95% confidence intervals.

	Reading skill model			Spelling skill model		
	Estimator	95% CI	p	Estimator	95% CI	p
(Intercept)	0.42	[-27.84, 28.69]	.976	11.01	[-5.19, 27.22]	.182
Age	3.76	[-0.65, 8.18]	.095	-0.02	[-2.55, 2.51]	.987
Gender	16.35	[-24.31, 57.01]	.430	2.52	[-20.79, 25.83]	.832
CST	6.04	[1.14, 10.93]	**.016**	0.72	[-2.08, 3.53]	.613
Person param.	1.88	[0.79, 2.96]	**<.001**	1.51	[0.89, 2.13]	**<.001**
Age:gender	-2.68	[-9.11, 3.75]	.413	-0.17	[-3.86, 3.51]	.926
Age:CST	-0.85	[-1.61, -0.08]	**.030**	-0.05	[-0.49, 0.38]	.808
Gender:CST	-1.00	[-8.07, 6.06]	.780	0.93	[-3.12, 4.98]	.652
Age:gender:CST	0.16	[-0.96, 1.27]	.783	-0.17	[-0.81, 0.47]	.599

low with values between $r^2 = 11 - 12\%$. When covariates are left out, r^2_{adj} drops down to about $4 - 8\%$. This finding reinforces the importance of covariates in model building, especially when the predictors alone are relatively low in predictive power. Overall, our SST results regarding influential predictors are consistent with the findings of Neubrand (2020). Unfortunately, the unusual response behavior of the children (e.g., constant selection of the same stimuli or constant switching of stimuli) could not be reliably detected and excluded accordingly due to a technical error in recording the children's input. In further analyses of the Rasch-models, Neubrand (2020) found, among other things, that the item difficulty did not increase throughout the task as it was intended when the material was created. Originally, items from higher categories were supposed to be more difficult. However, items from the first category with unequal numbers of syllable stress in the target and distractor had significantly higher guess probability and were not as difficult as intended compared to items from higher categories. An exception was item 12, whose difficulty was higher than the estimated person parameter of all children and therefore became too difficult. A detailed evaluation of the task can be found in the research of Neubrand (2020).

The sum score models and person parameter models are very similar in their performance if we keep in mind that models with BIC value within two units are practically equal (Burnham and Anderson, 2004). We finally decided to include the person parameter as a predictor of reading and spelling ability for further analysis. Due to its richer

Table 7.7: Overview of predictors resulting from SST analyses adopted for further machine learning analyses.

Literacy skill	Predictors
Reading	- person parameter
Spelling	- person parameter

informative value, e.g., accounting for guess probability per item, we find the person parameter more appropriate in terms of content. Although the explained variance is relatively low, we expect the inclusion of the person parameter from the SST in combination with predictors from other tasks to increase performance of the predictive literacy model. Regarding covariates, literacy prediction analysis of the Syllable Stress Task has encouraged us to consider CST and age in further analyses.

7.3 Rise Time Discrimination Task (RTDT)

As with the SST, we set up the following IRT models and compared them sequentially: Rasch model, Rasch model with guessing, 3PL model, 4PL model. A detailed description of these models can be seen in Table 7.8. Results, displayed in Table 7.8, show the 3PL model as the IRT model that best describes the data and the Rasch model with guessing as the most economical. When selecting an IRT model, we prefer the one that best fits the data and explains more variance (3PL) than decide based on its complexity level. From that 3PL model, we calculated the person parameter (pp), set up literacy prediction models as described in Table 6.1 and compared them in terms of prediction power. We found the correlation between the estimated person parameter and the sum score to be highly correlated, $r(359) = 0.97, p < .001$. We also set up sum statistic models and tested which is best suited for prediction. The results for both literacy skills can be seen in Table 7.9. We performed ANOVA tests to compare models with the sum score (ss), sum of clicks (soc) and their interaction for predicting reading and spelling skills. We did the same with the person parameter (pp) that substitutes for the sum score. No other model predicted literacy skills better than the simplest one, which only uses the sum score or the person parameter. This finding is persistent with and without covariates taken into account. However, models including the covariates have a much higher amount of explained variance (r^2_{adj}). When we compare the models with only the

Table 7.8: Chi-Square (χ^2) values and corresponding p-values to assess model fit, alongside model parsimony for the RTDT data. The 3PL showed the best and most fitting results but is not the most economical.

	χ^2	p	AIC/BIC
Rasch	-	-	9497/9578
Rasch with guessing	$\chi^2(20) = 38.12$	**<.01**	9498/9658
3PL	$\chi^2(19) = 31.66$	**.034**	9505/9738
4PL	$\chi^2(20) = 20.57$	.423	9524/9835

sum score as predictors to the models with only the person parameter, we see that their BIC is close to equal. Assuming the rule of thumb for the BIC, as suggested by Burnham and Anderson (2004), the models are practically equivalent for both literacy skills, with their differences being smaller than two units. Looking closely at the prediction model with the person parameter, we measured no significant contribution of the parameter, the covariates or their interactions. Estimators and confidence intervals can be seen in Table 7.10. These results don't change or improve when we look at parameter estimators for the sum score models.

7.3.1 RTDT discussion and summary

The comparison of models with total score and person parameters revealed no significant differences in performance. The sum score and individual characteristics did not appear to have any meaningful influence on predicting future language skills. Only when adding covariates did the explained variance increase to a notable level. However, the person parameter is a better choice if we want to include one of the two parameters as a predictor in our future analyses. Item-specific guessing probabilities are also considered when estimating the person parameter by the Rasch model with guessing. Thus, we can assume that the person parameter represents a purer performance measure in the Rise Time Discrimination Task than the sum score. In the results of Wölfl (2020), the intelligence measure and its interaction with age significantly contributed to the prediction of reading skills. We cannot reproduce this finding in our analysis. Apart from this, the results concerning task-specific performance measures such as the sum score and the person parameter are mainly similar. Generally, the results show that the constructed task does not contribute to explaining the reading and spelling data.

Table 7.9: The results of the RTDT data analysis show that there is no improvement when the sum of clicks (soc) or their interactions are added to the simple sum score (ss) or person parameter (pp) model. The comparison of the two top-performing models using the BIC score shows no difference in prediction power. Covariates include age, gender and intelligence.

		Models	Reading skill			Spelling skill		
			F-score	p	r^2_{adj}	F-score	p	r^2_{adj}
With	1.1	ss	-		8.38%	-		6.73%
covariates	1.2	ss+soc	$F(1) = 0.03$	.874	8.12%	$F(1) = 0.5$	.474	6.60%
	1.3	ss:soc	$F(1) = 0.06$	.800	7.87%	$F(1) = 1.18$	.279	6.64%
	2.1	pp	-		8.42%	-		6.64%
	2.2	pp+soc	$F(1) = 0.03$	.871	8.16%	$F(1) = 0.51$	.478	6.50%
	2.3	pp:soc	$F(1) = 0.07$	.792	7.94%	$F(1) = 1.59$	.208	6.67%
Without	3.1	ss	-		0.00%	-		0.12%
covariates	3.2	ss+soc	$F(1) = 0.00$	.951	0.00%	$F(1) = 0.57$	.450	0.03%
	3.3	ss:soc	$F(1) = 0.18$	.668	0.00%	$F(1) = 1.51$	.220	0.17%
	4.1	pp	-		4.27%	-		0.06%
	4.2	pp+soc	$F(1) = 0.00$	.948	4.04%	$F(1) = 0.57$	.452	0.00%
	4.3	pp:soc	$F(1) = 0.38$	.537	4.03%	$F(1) = 2.44$	.120	0.33%
Model	1.1	ss	$BIC = 2583$		8.38%	$BIC = 2200$		6.73%
comparison	2.1	pp	$BIC = 2583$		8.42%	$BIC = 2200$		6.64%

Table 7.10: Parameter estimators for the model with person parameter predicting spelling and reading skills, along with the 95% confidence intervals. If we substitute the person parameter with the sum score, we found no improvements or significant changes in the results.

	Reading skill model			Spelling skill model		
	Estimator	95% CI	p	Estimator	95% CI	p
(Intercept)	21.84	[-18.79, 62.47]	0.291	23.77	[0.17, 47.38]	0.048
Age	0.02	[-0.48, 0.52]	0.934	-0.16	[-0.46, 0.13]	0.270
Gender	14.93	[-43.60, 73.46]	0.616	4.42	[-29.59, 38.42]	0.799
CST	2.85	[-4.70, 10.41]	0.458	-2.18	[-6.57, 2.22]	0.331
Sum score	-0.41	[-1.48, 0.67]	0.456	0.26	[-0.36, 0.89]	0.407
Age:gender	-0.20	[-0.93, 0.54]	0.597	-0.04	[-0.47, 0.39]	0.847
Age:CST	-0.03	[-0.12, 0.07]	0.593	0.03	[-0.02, 0.09]	0.238
Gender:CST	0.75	[-10.34, 11.83]	0.894	2.20	[-4.24, 8.64]	0.502
Age:gender:CST	-0.01	[-0.15, 0.13]	0.908	-0.03	[-0.11, 0.05]	0.498

An in-depth analysis of this task by Wölfl (2020) shows further deficiencies in item and task creation and sequencing effects in the presentation order of items. For example, the model analysis showed that item difficulty was significantly higher for items where the distractor stimulus was presented first than for the target stimulus items. Further, the task might be too difficult as the difference in rise times seems too small to derive a true distinction between the groups. A deeper insight into the analyses, especially regarding the Rasch models and findings on item difficulties, can be found in the research of Wölfl (2020).

Despite these ambiguities in the RTDT and the moderately low r^2_{adj}, we would like to include the person parameter as a predictor. In combination with predictors from other tasks, we expect an additional value from the RTDT.

Table 7.11: Overview of predictors resulting from RTDT analyses adopted for further machine learning analyses.

Literacy skill	Predictors
Reading	- person parameter
Spelling	- person parameter

7.4 Serial Reaction Time Task (SRTT)

The first trial of each block was excluded because the experimental design incorrectly showed the first stimulus directly at the beginning of each block instead of showing the empty four boxes first. As a result, the response to the first trial in each block was delayed for most children. In addition, we excluded 1505 trials ($\approx 1.58\%$) the reaction times of which were more than three standard deviations away from the mean of the respective child to avoid data bias in the case of atypical trials of a child (Schmalz et al., 2019). Similarly, we excluded 884 trials ($\approx 0.94\%$), where the reaction time was below 100ms, as we assume that no intentional response is possible in such a short time.

The following analyses were calculated based on scenario 2. For reasons of space, the analysis based on scenario 1 has been moved to the appendix and is only mentioned in this chapter if there are significant deviations from scenario 2 results. If not mentioned separately, the results from the analysis of scenario 1 are consistent with the results from scenario 2.

7.4.1 Accuracy

A plot of the error rates across the blocks subdivided into the groups can be seen in Figure 7.2. It shows that children with weak reading and spelling skills are consistently less accurate than their counter peers across all blocks.

7.4.1.1 Results based on analysis by Lum et al. (2010)

The generalized mixed effects model with error rates as the target variable revealed a significant main effect of block ($Z = 11.35, p.001$) and group ($Z = 2.86, p = .004$) and a significant interaction effect of block and group ($Z = 2.66, p = .008$).

Except for blocks 1-2, 3-4, 3-5, and 4-5, all blocks were significantly different in terms of error rate, according to Bonferroni-corrected pairwise post-hoc tests. A detailed comparison of all blocks can be seen in Table B.7.

7.4.1.2 Results based on analysis by van der Kleij et al. (2019)

We checked the test conditions for the t-test for independent samples and found that no normal distribution can be assumed. Therefore, we computed a Wilcoxon rank sum test for independent variables and found that the groups are significantly different concerning

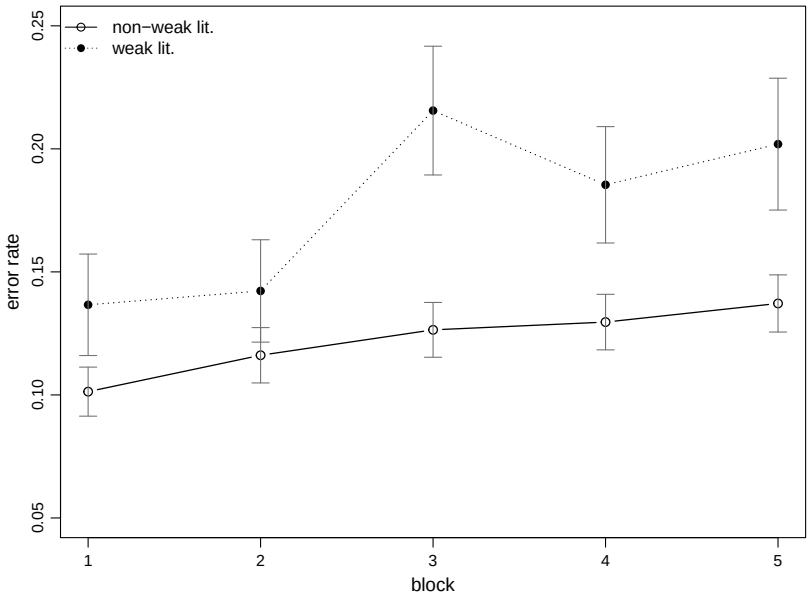


Figure 7.2: Progression of the groups' average error rates with standard error bars across the blocks. Children with weak literacy skills (weak lit.) appear to be less accurate in their responses, as their error rates are consistently higher than those of their peers (non-weak lit.).

their error rates ($W = 15939$, $p < .001$), with a mean of 87.86% accuracy for the children with non-weak literacy skills and 82.37% accuracy for children with weak literacy skills. For scenario 1 analysis, the respective mean accuracies were 85.60% and 90.74%.

7.4.1.3 Additional analysis by Metelmann (2020)

We calculated the Spearman rank correlation between error rates and mean reaction time for each child. The correlation was significant with $\rho = -0.4693$ ($S = 13757825$, $p < .001$). There was also a clear tendency for higher reaction times to be associated with lower error rates, as can be seen in Figure 7.3

7.4.2 Reaction Time

We followed the suggestion of Lum et al. (2010) for the reaction time analysis and excluded 11956 incorrectly solved trials ($\approx 13.07\%$) mainly to reduce unwanted bias. In the group with weak readers and writers, higher reaction times than in the control group were observed from blocks one to four. In both groups, reaction times were relatively consistent from blocks one to four. Both groups reacted slower in the fifth block (see

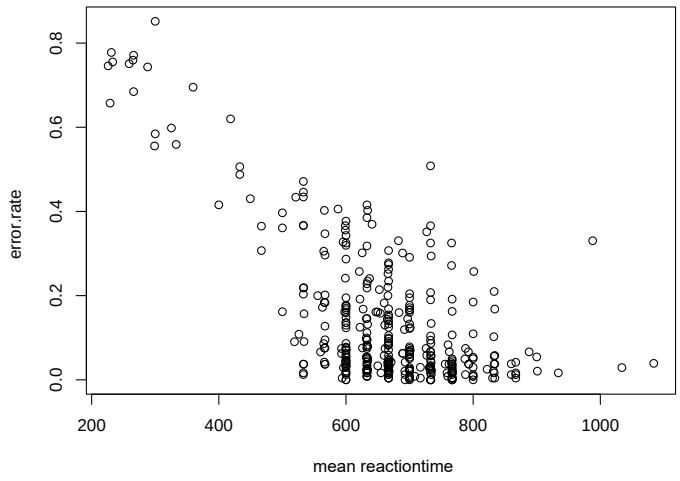


Figure 7.3: Mean reaction times in ms for each child compared to their error rates. A clear drop in the error rate with slower response times is evident.

Figure 7.4).

7.4.2.1 Results based on analysis by Lum et al. (2010)

As a basis for the evaluation of reaction times, we first calculated the reaction time differences between blocks 4 and 5 for each child. When analyzing variance with the reaction time difference as dependent and the group as an independent variable, the group was not significantly predictive of the reaction time increase from block four to five ($F(1) = 2.03,\ p = .155$). Looking at the variance analysis, the model including the reaction times of all blocks as the target variable and the group, the blocks and their interaction as independent variables, the blocks were significantly predictive of reaction time ($F(1) = 6.13,\ p < .014$). The group was neither predictive of reaction times in the main effect ($F(1) = 0.03,\ p = .860$) nor in the interaction with block ($F(1) = 0.289,\ p = .590$). To interpret the block's main effect, we further compared the mean reaction times of all blocks in pairs using a Wilcoxon rank sum test for dependent samples. A Bonferroni correction was applied. Results and Figure 7.4 show that blocks one to four were all significantly different from block five (for more detail, see Table B.8). Since a log transformation, as Lum et al. (2010) used in their analysis, did not lead to a normal distribution of the reaction time variable, we did not transform the data and

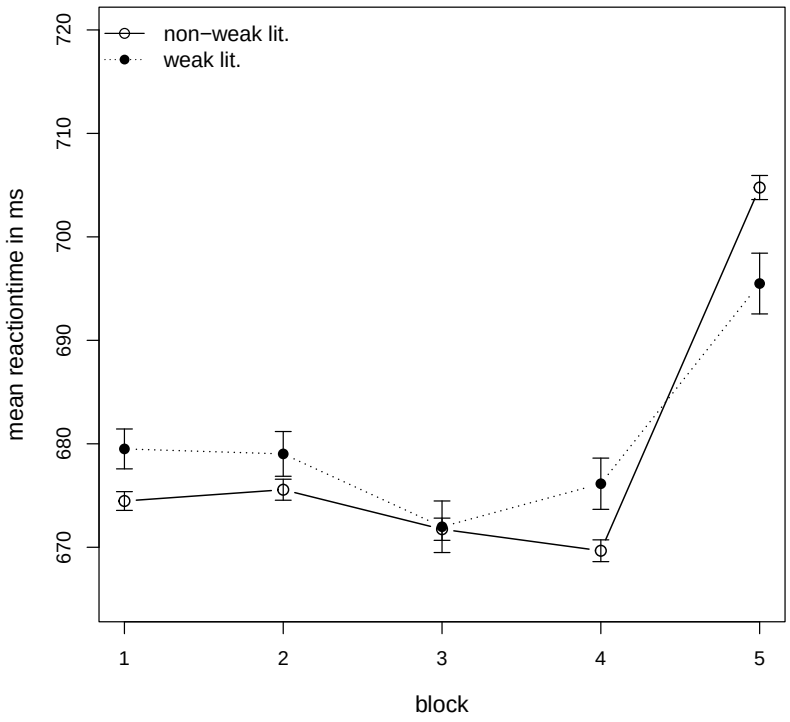


Figure 7.4: Mean reaction times in ms per group and block with standard error bars. Children with weak literacy skills appear to be slightly slower in learning blocks one through four and appear to have a smaller drop-off to block five than their peers.

used a Wilcoxon test instead of a t-test for each block comparison.

7.4.2.2 Results based on analysis by van der Kleij et al. (2019) and Metelmann (2020)

We fitted piecewise growth models (using *lme4*-package for R) starting from the most simple model with a linear learning curve, response increase, and group as fixed effects and learning curve and response increase as random effects ($BIC = 22313$). Only the response increase contributed significantly to the prediction of reaction times ($t(382) = 6.88,\ p < .001,\ BIC =$), whereas group ($t(381) = 0.06,\ p = .954,\ BIC =$) and the linear learning curve ($t(382) = -1.11,\ p = .269,\ BIC =$) did not. Adding interactions of these fixed effects did not improve model prediction, as the variance analysis showed ($\chi^2(2) = 1.53,\ p = .466,\ BIC = 22326$). Replacing the linear term

of the learning curve to a quadratic one did not increase the prediction performance ($\chi^2(3) = 1.60,\ p = .658,\ BIC = 22334$). The likelihood-ratio test when adding a random slope with quadratic term did get significant compared to the linear learning curve model ($\chi^2(7) = 33.65,\ p < .001,\ BIC = 22332$) as did the model that added a cubic term with a random slope with cubic term ($\chi^2(7) = 1.53,\ p =< .001,\ BIC = 22328$), but the BIC is distinct higher. Despite the test result, we prefer the simpler model since we predominantly rely on the BIC criterion for making model selection decisions. Adding a sole cubic term did not increase predictive power ($\chi^2(3) = 37.26,\ p = .670,\ BIC = 22334$).

7.4.3 Results based on further analysis by Metelmann (2020)

With a linear learning curve and response increase as fixed effects and a learning curve and response increase as random effects, we first fitted piecewise growth models ($BIC = 22304.99$). Response increase contributed significantly to the prediction of reaction times ($t(383) = 6.89,\ p < .001$), whereas the linear learning curve ($t(383) = -1.11,\ p = .268$) did not. Adding a quadratic term to the learning curve did not contribute significantly to model prediction, as the BIC comparison shows ($BIC = 22312.47$). Adding a cubic term to the learning curve instead of a quadratic did not contribute significantly to model prediction ($BIC = 22312.52$).

7.4.3.1 Prediction model for reading

We calculated the individual learning curve, response increase and intercept coefficient for each child from the previously established, most fitting reaction time model. We set up and fitted the reading prediction model with those parameters as predictors. An overview of the model and predictor performance after each stepwise reduction can be seen in Table 7.12. In the first step, we identified the intercept coefficient as the predictor with the least evidence, removed it from the model and compared the newly fitted model with the more complex one in the second step. All remaining predictors significantly impacted explaining the data, so we did not reduce the model further. The model that included accuracy, a quadratic learning curve, and response increase as predictors achieved the best fit for predicting reading ability.

Table 7.12: Stepwise reduction of reading prediction model for the SRT task, starting with the most complex model. In each step, the predictor with the least evidence (p) is dropped until only significant predictors are left.

Predictors	1. Step			2. Step		
	t value	p	drop	t value	p	drop
Accuracy	3.17	**.002**		3.40	**<.001**	
CoefIntercept	-0.01	.991	x			
CoefBlockLearn	-2.20	**.028**		-2.41	**.016**	
CoefRespIncrease	2.25	**.025**		3.39	**<.001**	

7.4.3.2 Prediction model for spelling

For the spelling prediction model, we set up and fitted the model with accuracy, intercept coefficient, learning curve coefficient and response increase coefficient generated from the reaction time model as predictors. In the first step, we identified the intercept coefficient as the predictor with the least evidence, removed it from the model and compared the newly fitted model with the more complex one in the second step. All remaining predictors in the reduced model seem to have a significant influence in describing the data. We therefore don't need to further reduce the model. An overview of the spelling models and the results of the stepwise reduction can be seen in Table 7.13.

Table 7.13: Stepwise reduction of spelling prediction model for the RAN task, starting with the most complex model. In each step, the predictor with the least evidence (p) is dropped until only significant predictors are left.

Predictors	1. Step			2. Step		
	t value	p	drop	t value	p	drop
Accuracy	3.63	**<.001**		3.95	**<.001**	
CoefIntercept	0.16	.874	x			
CoefBlockLearn	-2.24	**.026**		-2.38	**.018**	
CoefRespIncrease	1.45	.149		1.99	**.047**	

7.4.4 SRTT discussion and summary

Accuracy seems to be a good indicator to distinguish between groups across both literacy skills, as results and Figure 7.2 show. Therefore, the accuracy score is added as a predictor for future analysis.

When looking at our reaction time analysis of the evaluations similar to Lum et al. (2010) and van der Kleij et al. (2019), the group did not contribute in a meaningful way to the prediction of the data neither as a main effect nor in interaction with the learning curve or the response increase. The group-specific reaction time trend expected by Lum et al. (2010) is neither evident across learning blocks nor between learning blocks and the pseudorandomized block. There seems to be no learning in both groups over the first four blocks, and there seems to be an increase in reaction time, even for the non-weak literacy group in the last block.

Since implicit learning through constant repetition of the same sequence is not present in blocks one to four, and we even see a trend towards slower reactions, we took a closer look at the course of the reaction times within the blocks, see Figure 7.5. As can be seen, the reaction times increased from sequence to sequence within the blocks. This development gives rise to the assumption that there is a kind of fatigue effect on the children during the exercise. A comparison between the reaction time of the first sequence in block one and the first sequence in block five reinforced this assumption. Both unique sequences were presented to the children for the first time at this point, which is why the same conditions could theoretically be assumed. However, the response times in the first sequence in block five was significantly higher than the first sequence in block one when calculating a Wilcoxon rank sum test with continuity correction ($W = 64903,\ p = .006$).

As can be seen from Figure 7.5, the behavior patterns in block two to four appear to be very similar. This finding can be used to reduce the complexity of the task and thus counteract the fatigue effect. For example, reducing the number of sequences from five to three in each block or deleting the third and fourth blocks would most likely still produce the same trends but reduce task complexity.

When predicting literacy skills using the generated parameters from reaction time analyses, the individual learning block coefficient and the response increase coefficient helped significantly explain the data for both literacy skills. Therefore, besides accuracy, we

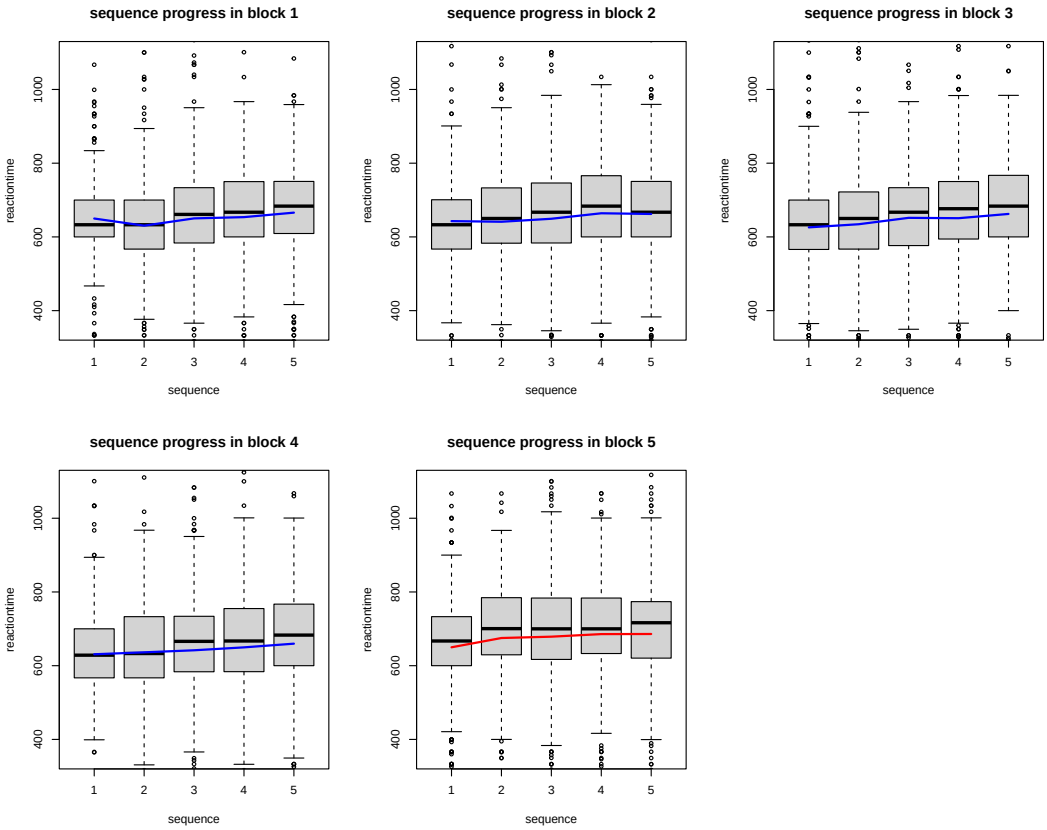


Figure 7.5: Reaction time progression of the sequences across all blocks. It can be seen that children respond more and more slowly within a block and that this pattern continues, at least with learning blocks 1-4. The mean reaction time per sequence is indicated by the blue and red lines, respectively.

added these coefficients to our set of predictors for future analysis.

7.5 Rapid Automatized Naming Task (RAN)

The linear model confirmed that the order of the pages shown had no effect on children's performance on the task. This means, that the second page of stimuli children are asked to name, does not slow down overall naming time or introduce any unwanted bias. Page as an independent variable was not predictive of the itemtime[16] ($F(690) = 1.05$, $p = .306$).

As Figure 7.6 shows and a Wilcoxon rank sum test with continuity correction for the

[16] *Itemtime* refers to the average time it takes a child to label an item.

Table 7.14: Overview of predictors resulting from SRTT analyses adopted for further machine learning analyses.

Literacy skill	Predictors
Reading	- accuracy
	- learning curve coefficient
	- response increase coefficient
Spelling	- accuracy
	- learning curve coefficient
	- response increase coefficient

non-normal distributed itemtime confirmed, we found that the groups are significantly different with respect to their itemtimes ($W = 54182$, $p < .001$), with a mean of 1.18s per item for the non-weak reader and writers and 1.36s for the weak reader and writers.

7.5.1 Prediction model for reading

Since there was no sequence effect, we calculated the individual itemtime, amount of wrong-naming errors and amount of omitted-words errors for all items across both pages. We fitted the reading prediction model with those parameters as predictors. In the first step, we identified the omitted-words error as the predictor with least evidence, removed it in the second step and fitted the new model. Wrong-naming did not contribute significantly to reading skill prediction in the second step, which left itemtime as the only predictor of importance. An overview of the reading models and the results can be seen in Table 7.15. The same analysis with covariates age, gender and CST showed that CST made a significant contribution in explaining the data in the model with only itemtime as predictor. In order not to go beyond the scope of the analysis here, we moved the listing of the analysis with covariates to the appendix in Table B.9.

7.5.2 Prediction model for spelling

For the spelling prediction model, we set up and fitted the model with overall itemtime, wrong-naming error and omitted-words error of each child as predictors. The most complex model contained all predictors. We reduced the model by the predictor with the least evidence, omitted-word error. The remaining two predictors form the model in

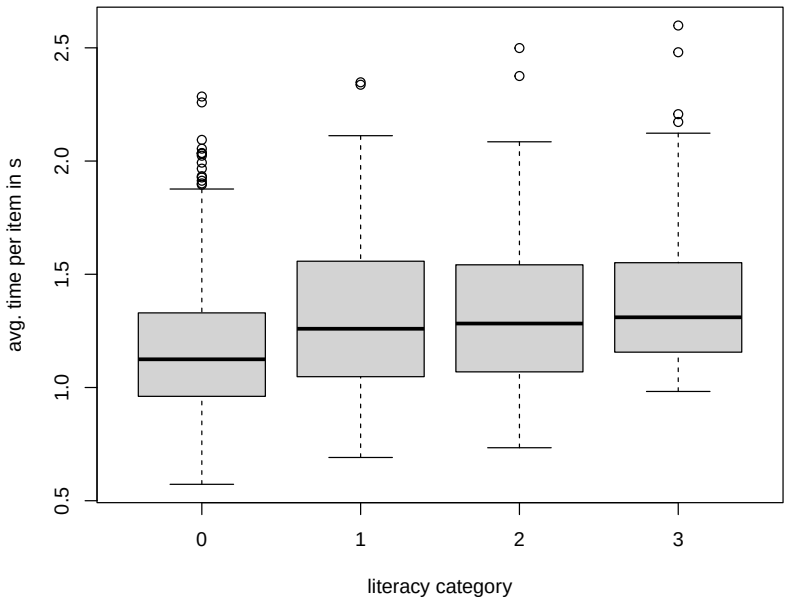


Figure 7.6: The average itemtime across all items for each of the four literacy skill groups. Literacy category 0 refers to children we considered to have no reading and spelling difficulties, 1 to children with only reading difficulties, 2 to children with only spelling difficulties and 3 to children that we considered to have difficulties in both skills.

the second step and both contributed significantly in predicting the spelling skill. We therefore stopped further model reduction. An overview of the spelling models and the results can be seen in Table 7.16. These results and the order of predictors removed are maintained even if we include covariates age, gender and CST from the same analysis. Therefore these results from models with covariates are not listed.

7.5.3 RAN discussion and summary

The time taken by the children per item seems to be a good measure to distinguish between the weak and non-weak literacy children. On average, weak children were consistently slower at naming. The amount of wrong-naming errors as an additional predictor for spelling skill prediction significantly improved prediction performance. Age, gender or CST do not appear to play a role in predicting spelling skills with the RAN task, whereas the CST showed significant influence for reading skill prediction.

Table 7.15: Stepwise reduction of reading prediction model for the RAN task, starting with the most complex model. In each step, the predictor with the least evidence (p) was dropped until only the predictor itemtime, which contributes significantly in reading skill prediction, was left.

Predictors	1. Step			2. Step			3. Step		
	t value	p	drop	t value	p	drop	t value	p	drop
Itemtime	-6.62	**<.001**		-6.67	**<.001**		-6.67	**<.001**	
Wrong-naming	-0.79	.432		-0.89	.377	x			
Omitted-word	-0.78	.433	x						

Table 7.16: Stepwise reduction of spelling prediction model for the RAN task, starting with the most complex model. In the first step, the predictor omitted-word has least evidence (p) and is dropped. The remaining predictors itemtime and omitted-word contribute significantly in spelling skill prediction, which is why stepwise reduction ends.

Predictors	1. Step			2. Step		
	t value	p	drop	t value	p	drop
Itemtime	-5.00	**<.001**		-4.88	**<.001**	
Wrong-naming	-2.32	**.02**		-2.20	**.03**	
Omitted-word	1.10	.274	x			

Effects, such as fatigue or sequence effects, which seemed to have had a greater or lesser influence on the other tasks, did not influence the RAN. Children had steady performance throughout the task.

7.6 Importance of covariates

In none of the tasks did gender have an effect on predicting literacy skills. We did not consider this covariate for further analyses based on this finding. The intelligence measure CST and age significantly influenced the literacy prediction models of the RAN task and the SST. Their interaction also significantly impacted predictions, which makes sense given that an older child is expected to be more intelligent than a younger child. We will include age and CST as covariates because we want to account for the influence of external variables that could affect the results of the planned machine learning model

Table 7.17: Overview of predictors resulting from RAN Task analyses adopted for further machine learning analyses.

Literacy skill	Predictors
Reading	- time per item in s (itemtime)
Spelling	- time per item in s (itemtime)
	- wrong-naming error count

analysis.

7.7 Summary of task results

After performing statistical analysis and evaluation, we looked at how well each task predicted reading and writing proficiency. We then analyzed and evaluated their predictive features, answering *RQ1.4*. The result of these efforts is presented in Table 7.18, which presents a summary of all predictors considered for our machine learning analysis.

Although some screening tasks may not directly measure literacy skills, incorporating their performance measures as predictors can enhance the overall generalization of the literacy screening. To support this claim, we ran a Principal Component Analysis (PCA) (Pearson, 1901) on the predictors listed in Table 7.18 as well as age, CST and whether German was stated as native language as covariates. With the partial correlation as input, we were able to reduce the eleven predictors to five relevant components - or *dimensions* - with the Kaiser-Guttman criterion for component's eigenvalues greater than 1 applied. Those components are linear combinations of the original predictors that explain the most variance in our data. 3D plotting the loadings of these components helps us visualize how our original predictors are projected onto the first three principal components. Figure 7.7 shows us that the predictors are widely scattered across these three dimensions. As a result, we can assume that the predictors, and thus our tasks, measure and cover a broad spectrum of proficiencies.

Table 7.18: Summary of predictors derived from the tasks that will be used for machine learning analysis.

	Literacy skills	
Task	Reading	Spelling
IPET	• response bias c_{real}	• response bias c_{real}
	• response bias $c_{fictitious}$	• response bias $c_{fictitious}$
	• performance indicator d'_{real}	• performance indicator d'_{real}
	• performance indicator $d'_{fictitious}$	• performance indicator $d'_{fictitious}$
SST	• person parameter	• person parameter
RTDT	• person parameter	• person parameter
SRTT	• accuracy	• accuracy
	• learning curve coefficient	• learning curve coefficient
	• response increase coefficient	• response increase coefficient
RAN	• time per item	• time per item
		• wrong-naming error count

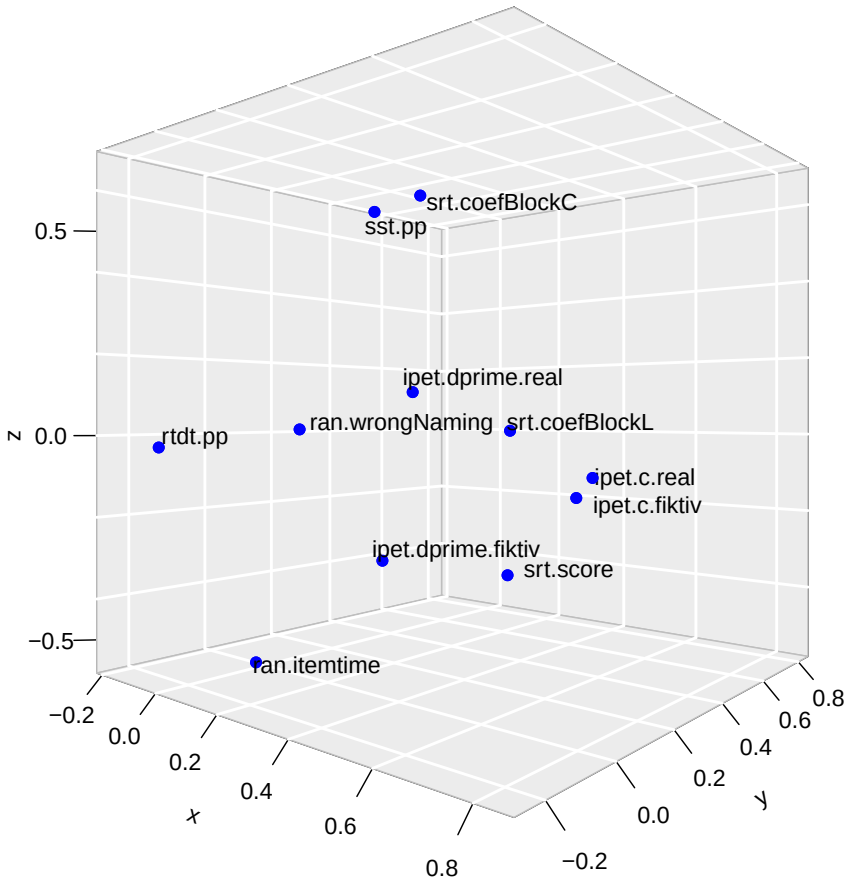


Figure 7.7: Visualization of the predictor loadings in a 3D space, representing the top three dimensions, after Principle Component Analysis.

Part C

Predictive Data Analysis

Chapter 8

Machine Learning and its Application in the Literacy Screening

Machine learning is now a well-established technique for extracting information from and understanding complex datasets. In this chapter, we incorporate machine learning approaches into a literacy screening environment, explaining the key ideas to the extent required to follow the arguments and replicate the approach.

Our screening project consists of two target variables we aim to predict as accurately as possible: reading skill (as measured by test scores from a standardized reading test) and spelling skill (as measured by test scores from a standardized writing test). For each of these skills, separate prediction models are required. This presents the challenge of handling the dataset for each model type differently and setting up the model training and evaluation in such a way that a separate evaluation of each model is possible, as well as a joint evaluation to compare our screening as a whole with other screenings.

As is known and common with dyslexia screenings (e.g., in the research of Rauschenberger et al. (2020b)), the underlying data often have unique characteristics that complicate the preparation and processing of the data for model building and model evaluation. Small sample sizes, skewed distribution (resulting in a skewed model), and multiple predictors with unknown influences introduce significant challenges to developing models that accurately predict children's reading and spelling performances. In this chapter, we first introduce the relevant basics of machine learning and then derive parameters and approaches for our application that will help overcome the aforementioned challenges.

8.1 Developing the machine learning task

Literacy screenings aim to detect children's literacy weaknesses and are used to make predictions about children's reading and spelling skills. Training learning algorithms can generate predictions based on various input parameters, e.g., multiple test performance indicators. These predictions can be discrete or continuous values depending on the desired outcome. In the following paragraphs, we outline the two basic predictive approaches of regression and classification in the context of literacy screenings to motivate the reasoning for choosing the approach adopted in this study.

Regression algorithms are mainly used to predict a continuous target variable (e.g., age, price, income, score), whereas classification algorithms aim to predict a discrete target variable (e.g., male/female, spam/no-spam, dyslexic/non-dyslexic). Deciding which algorithm is best for literacy screenings and the dataset at hand depends on how you want to present the result. In most cases, the goal of dyslexia or literacy screening (see also related research discussed in Section 2.2) is to classify a child as *at-risk* or *not at-risk* for reading and spelling. This classification is achieved by measuring children's performance and applying a threshold for subsequent categorization of the collected results. While discretizing a continuous variable may be necessary, for example, to set a cutoff value for diagnostic criteria, it is usually not a good idea because information is lost. Dividing a continuous value may lead to the loss of power when doing hypothesis tests (Van Belle, 2008), to a difficult interpretation of the results, especially when the domain is changing over time (Good and Hardin, 2008), and allows for easy (and potentially malicious) alteration of the data (Wainer et al., 2006).

Classifications in the context of literacy screenings allow a clear statement as to whether the child is *at-risk* of dyslexia and further make those screenings somewhat comparable to each other. However, it is often unclear on what basis the results are classified or how certain the given result is. Even current German analog screenings often include a third, edge-case class to categorize children that are somewhere between both classes (Marx and Lenhard, 2011). When using regression, this loss of information or chance of total misclassification can be reduced or even avoided by, for example, predicting the reading score of a child and including uncertainty rather than strictly categorizing their performance.

As mentioned before, our screening aims at predicting literacy skills presented as continuous values rather than binary classes. Therefore, we decided to formulate our task as

a regression task with the goal of predicting the continuous reading and writing scores. To the best of our knowledge, none of the aforementioned screenings mentioned in Section 2.2 attempt to predict any particular literacy skill but instead try to classify the literacy of given children according to the test results.

8.2 Handling an imbalanced dataset

Real-life applications and screenings of all sorts often struggle to get accurate predictions due to their data being skewed towards one class or value of the target variable. Dealing with these imbalances is an integral part of data preparation prior to training models and a challenging task when working in a regression environment. In this section, we first take a look at why it is important to address imbalances in a dataset, why it is more complicated in regression and what options are available to deal with these problems. In order to optimally prepare our data, we finally explore the extent of imbalance present in our data and how we can deal with it for model training.

8.2.1 Definition of imbalanced dataset

An imbalanced dataset usually refers to learning data in a classification setting where at least one class is under- or over-represented. Detecting fraud, meteorological disasters, or medical diagnoses are just a few real-world examples where the more meaningful outcome is less likely but still as important to predict. Such problems become evident during the training of classifiers when the asymmetry of misclassification costs leads to an undesirable bias in the model prediction. For example, this can lead to the classifier assigning every entry in the test set to the majority class because that's where it obtains the best accuracy score. This problem has thoroughly been explored in classification, but similar issues are present for regression problems. Instead of the underrepresentation of classes, values in a relevant, mostly extreme region of a spectrum are rare and therefore poorly represented in the training data. As mentioned in Section 8.1, it is possible to classify a continuous variable and, for example, convert different ages into different classes. However, this approach leads to a decreased comparability between individuals of the same age. In both cases, i.e., classification and regression, the predictive performances are disappointing if the imbalanced training data are not treated.

8.2.2 Handling imbalanced datasets in regression

Generally, there are two main strategies to handle imbalances in a dataset: processing the data or processing the learning algorithm. While data processing can occur before (preprocessing) and after (postprocessing) an algorithm is applied, model processing changes the algorithm. Postprocessing data involves adjusting a model's predictions after the fact, which often results in a loss of interpretability of the model (Torgo, 2017). This approach is not useful for our application as we extensively analyze the input data and want to interpret the outcome based on the used features. Therefore, in this section, we focus on data preprocessing and model processing for handling imbalances.

8.2.2.1 Preprocessing the data

The goal of data preprocessing is to transform the training data before applying any learning algorithm to it. This can be achieved by changing the distribution (sampling) of the target value or changing their weights (weighting) when learning from it. The main advantage of both approaches is that the choice of learning algorithm for model training is not restricted in either.

Sampling The sampling strategy is one of the most successful approaches to deal with imbalances (Branco et al., 2017). However, it is unclear what the perfect distribution for target training data should look like and how much alteration should be applied. Oversampling and Undersampling are well-established sampling preprocessing strategies. Both methods can be applied together or separately until target training distribution is achieved. The number of cases in the affected underrepresented areas is increased in oversampling. In its simplest form, this transformation can be achieved by duplicating relevant data points. However, with this approach, the probability of perfectly fitting the predictive model only on the training data (overfitting) is high due to its lack of data heterogeneity(Fernández et al., 2018). More sophisticated algorithms should be applied, e.g., Synthetic Minority Oversampling Technique (SMOTE) for regression or Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN). These algorithms use functions like k-nearest neighbor to generate new, synthetic data points as shown in Figure 8.1.

In undersampling, one reduces the number of cases in the overrepresented areas. Removing large amounts of data can make it difficult for the learning algorithm to distinguish

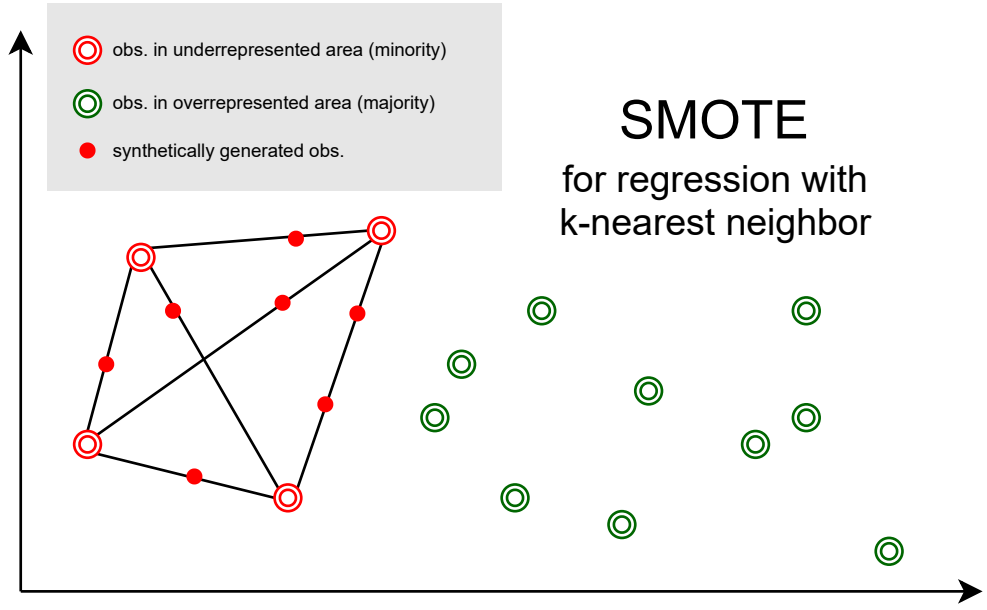


Figure 8.1: The Synthetic Minority Oversampling Technique (SMOTE) algorithm for oversampling is mostly used in classification use cases but can also be transformed into a regression setting. It uses the k-nearest neighbor algorithm to generate new data based on the surrounding observations.

between relevant and irrelevant values (He and Ma, 2013). Due to this limitation and the fact that this procedure is mainly recommended for larger datasets, we decided not to further consider this option for our screening data.

Weighting The data distribution remains unchanged in weighting, but the relevant observations are given weights. This is a commonly used and easy-to-implement method in classification but is yet to see popular usage in regression due to its complicated application with continuous values. One challenge is that the learning algorithm needs to understand and incorporate the weights in its calculations, e.g., in the cost functions. With most standard algorithms, this needs to be implemented manually and is not known to produce significantly better results.

8.2.2.2 Processing the learning algorithm

Another way to deal with unbalanced data is to adapt existing algorithms to one's needs or to use algorithms that are capable of dealing with unbalanced data in the first place.

Robust Algorithms Some algorithms are inherently more robust to imbalanced data than others. For example, tree-based models such as decision trees attempt to partition heterogeneous data into homogeneous subsets to increase the purity of observations. Suppose the distribution of the variable to predict is unbalanced. In that case, the model will likely learn splits at the beginning of its algorithm, separating the majority from the minority areas. This assumes that the minority data all lie in one area of the feature space. When learning with an ensemble of decision trees, this process becomes even more robust to imbalances.

Model Processing A more complex approach to handling imbalanced data is model processing. Promising techniques like utility-based regression (UBR), as proposed by Torgo and Ribeiro (2007), seems to work efficiently in many data mining application. Due to the mostly unexplored nature and the challenging implementation, which, among other things, requires a custom cost penalty with the use of relevance functions for each model tested, we did not consider this approach for the project in the described work.

8.2.3 Handling data imbalances in the screening dataset

A look at the screening data shows that in both target variables, the important areas of low-performing children are underrepresented (Figure 8.2). As mentioned in Section 2.2, all children with reading or writing scores that are less than one standard deviation below their mean are considered low-performing or weak. A regression model trained on the dataset shown in Figure 8.2 would be less likely to predict future underperformances of children.

Considering the skewed structure of the dataset and the possibilities mentioned in the previous section, we decided on a mixture of data preprocessing and algorithm processing methods to tackle the issue of imbalances in our screening dataset.

- *Favor sampling methods over weighting.* When working with continuous target variables, we found that sampling methods were more beneficial and easier to use. With sampling, we could use most standard algorithms out of the box, whereas with weighting, we would have had to use specialized or customized models. Additionally, the implementation of sampling methods seemed to be more supported in the R-framework *caret* than weighting when using regression.

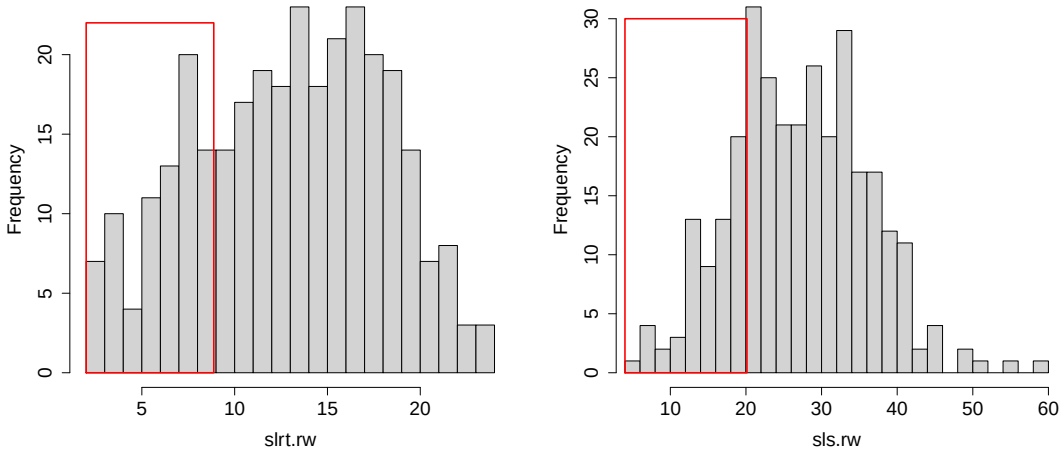


Figure 8.2: Histogram of the SLRT (left side) and SLS (right side) raw values (n=306). Marked in red are all low-performing children with raw values below 1SD of the mean.

- *Oversample only data of interest.* When training the models, we added synthetically generated new observations for low-performing children to the training set to increase their importance in the model representation (oversampling). The counterpart undersampling[17] would further reduce our sample size for the present dataset and lower the generalizability of our models.
- *Make use of established learning algorithms.* To find the best model, we must first identify regression algorithm models applicable to our use cases. Further research and similar screening projects (e.g., Rauschenberger et al. (2022)) provided us with a good starting point for determining which models to consider.
- *Use comparable learning algorithm models.* We decided against model-modifying processing due to its complexity in implementation and interpretation. Using multiple frameworks and implementations of these models while enabling model processing would also further limit interpretability and make comparisons between models difficult. A common ground for comparison can easily be achieved by using the same framework for all models.

Given these techniques and methods, we plan to address *RQ 2.1* on how to handle

[17] Undersampling is a technique for combating a skewed distribution by removing samples from the overrepresented area.

imbalances in our dataset.

8.3 Training and evaluating a model

In this section, we explore the model training and evaluation process. We look at the typical modeling workflow and further investigate different subsampling techniques that help generate better and more robust training and evaluation results. We also compare and decide on a cross-validation method using results from early pre-tests.

To avoid confusion in the naming of different datasets used when training and evaluating, the terms are briefly summarized in Table 8.1.

Table 8.1: Differentiation of the terms used in model training.

Term	Description
learning set	the complete dataset available to build and evaluate a model
training set	generated from the learning set; used for further model training
test set	generated from the learning set; used to evaluated the finalized model on
validation set	generated from the training set as part of the cross-validation; functions as a test set within cross-validation

8.3.1 Modeling workflow

Figure 8.3 shows a typical workflow on how a learning set is processed to be used for model training and evaluation. This approach is widely used in machine learning(Pedregosa et al., 2011) and is also the basic approach for our screening analysis.

When evaluating the model, subsampling techniques are used to split between training and test sets. Cross-validation techniques are frequently used to optimize model training and determine the best model parameters. In the next sections, we will take a closer look at both techniques.

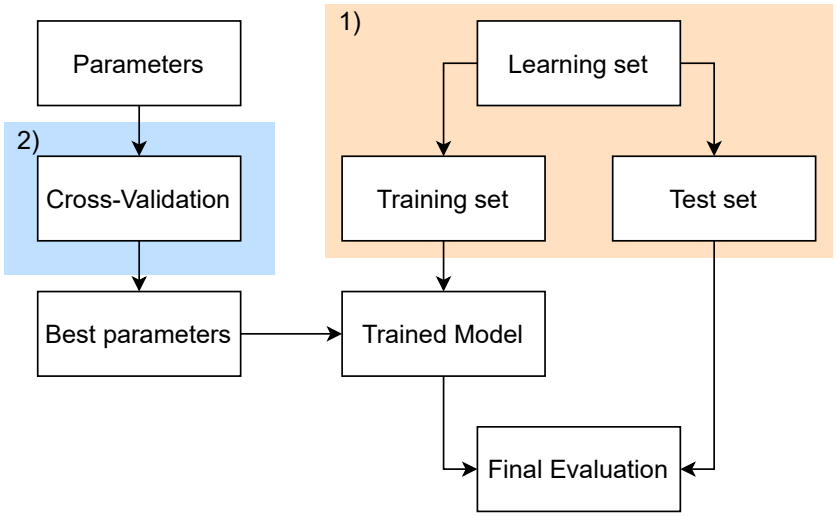


Figure 8.3: The general workflow for model training and evaluation. In section 1) single hold-out random subsampling is used to split the learning set into a training and a test set. In section 2) cross-validation methods are used to optimize model training and determine the best parameters.

8.3.2 Subsampling

Subsampling, particularly cross-validation subsampling, is an important technique for generalizing predictive models and avoiding overfitting in models (Berrar, 2018). The basic idea behind subsampling is to run the learning algorithm only on a subset of the data (training set). The remaining subset (test set) is then used to evaluate the trained model. This ensures that the model is evaluated using valid and previously unseen data from the same population from which it was created. The following descriptions of sampling methods are summarized from Berrar (2018).

8.3.2.1 Single hold-out random subsampling

The most common and simplest subsampling strategy is the single holdout technique, where cases are randomly selected from the learning set for the test set, while the remaining cases form the training set. 70% to 90% of the cases are commonly used for the training set and 30% to 10% of the cases for the test set, mainly depending on the size of the dataset. When randomly splitting data in this way, it is often beneficial to ensure that the distribution of the original dataset is reflected in the newly split data. This method is called stratified sampling. The training set in a stratified single-holdout random subsample would have a similar distribution of data as the learning set from

which it was created. In the frameworks we researched, such as *caret* and *scikitlearn*, only data-split methods are provided that are based on the distribution of a single target variable.

8.3.2.2 Custom stratified subsampling method for two variables

In order to evaluate the models on a representative dataset, it is important, especially to see how they perform in an imbalanced dataset, to preserve the original distribution of each target variable as much as possible when splitting into training and test data. In addition, we need to be able to predict both target variables for each child in the test dataset to examine the prediction results of reading and writing performance separately and in combination. This will also be needed to compare the performance of our literacy screening with other dyslexic/literacy screenings.

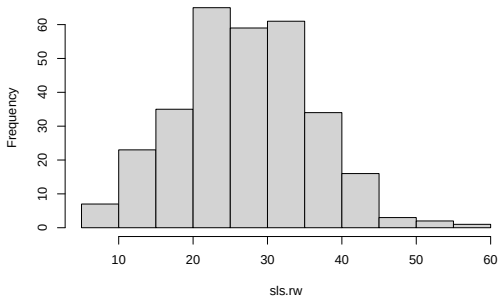
The aforementioned specifications lead us to a two-variable stratified, single-hold-out, random subsampling algorithm that splits the learning set (Figure 8.4a) into a training set (Figure 8.4b) and a test set (Figure 8.4c). We decided on a split ratio of 70% training and 30% test set so that we have a relatively large test set for optimal evaluation of the models. A custom implementation is inevitable because the R-framework *caret* doesn't provide a two-variable stratified subsampling method.

The procedure for our custom two-variable stratified, single-hold-out, random subsampling algorithm is implemented as follows:

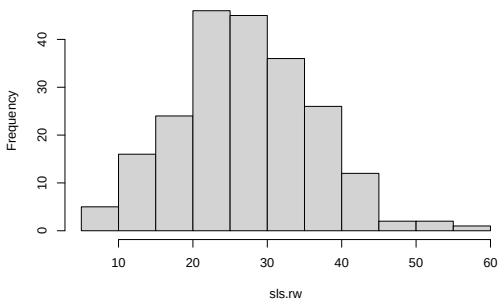
- Binning both target variables independently using the stanine scaling method
- Merging both datasets
- Grouping the dataset by both bins with randomness included if required
- Sample the desired fraction of the train set
- Extract the test set by anti-joining the train set sample

8.3.3 Cross-validation

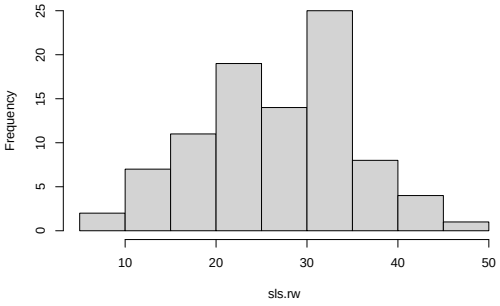
Overfitting is a methodological mistake that can occur when learning and assessing the parameters of a prediction function against the same data. A model replicating the labels of the samples it has just seen would receive a perfect score but would not generalize well on unseen data and fail to predict anything meaningful. To avoid this, resampling methods, such as cross-validation, are used (Berrar et al., 2013).



(a) Distribution of reading raw value on the complete dataset.



(b) Distribution of reading raw value on the 70% training set.



(c) Distribution of reading raw value on the 30% test set.

Figure 8.4: Stratified splitting of the SLS 2-9 raw value collected in tests into a training and test set.

8.3.3.1 K-fold cross-validation

In k-fold cross-validation, the learning set gets partitioned into k-equal and disjoint subsets (folds). Cases are selected at random and without replacement for each fold. The model is then trained with $k-1$ folds and applied to the remaining fold, called the validation set, to measure the performance of the model. This method is continued until all k subsets have been used as validation sets. The cross-validated performance is the average of the k performance measurements on the k validation sets. For real-world datasets Kohavi (1995) recommend stratified k-fold cross-validation. A reflection of the original distribution in the data subsets helps to avoid a biased evaluation.

K-fold cross-validation is sometimes repeated r times with different k-fold subsets to minimize the variance of the predicted performance metric. A cross-validation with 10-fold

that is repeated five times is called five times repeated 10-fold cross-validation. However, Molinaro et al. (2005) found that such repeats only modestly reduce variance. As a result, the effort required to implement repeated cross-validation must be proportional to the benefits of the method.

8.3.3.2 Leave-one-out cross-validation

When the amount of folds is set equal to the total observations ($k = n$), this technique is a particular instance of k-fold cross-validation called leave-one-out cross-validation (LOOCV). In this scenario, each individual case acts as a hold-out case for the validation set. For example, in the first iteration, the first case x_1 serves as the validation set, and the remaining cases x_2 until x_n serve as the training set. This process is repeated until each case has served as a validation set once. LOOCV's computational cost can be significant for large n due to the high number of folds and iterations this method must evaluate.

8.3.3.3 Comparing k-fold and leave-one-out cross-validation

During the development phase, preliminary tests revealed that there are no substantial differences in model performances when k-fold cross-validation or LOO cross-validation is used. The resulting model performances did not significantly differentiate between models trained with 10-fold cross-validation and LOOCV ($p > 0.9$). As a result, and due to the slight increase in computational resource requirements, LOO cross-validation was no longer considered, and we trained all models using k-fold cross-validation. We did not include repetition because we needed to implement custom k-fold cross-validation (see the following section). The implementation effort seemed to be greater than the benefits, as research of Molinaro et al. (2005) showed.

8.3.4 Integrating oversampling into model training

In Section 8.2.2, we discussed oversampling as a method for preprocessing our imbalanced screening data. If we want to integrate oversampling into our training process, we need to answer the following questions first:

1. Is oversampling beneficial for predicting low-performing children, specifically in our case, and if so, how high should the oversampling rate be?

2. In conjunction with cross-validation to improve model generalization, at what point in the training process is oversampling appropriate?

Regarding question 1: A system that allows us to test and analyze multiple oversampling percentages for all relevant models is required to answer this question. With the integration of this system into our model training process, we generate models trained on a non-oversampled training set and models trained on oversampled training sets.

In the first analysis, we compare the general performance during training between the models trained on non-oversampled versus oversampled datasets. We expect that the addition of synthetically generated data to the training set will most likely have an impact on the models' performance. Depending on the level of oversampling, it is possible that a new imbalance in the distribution of the data will occur, such that the observations of underperforming children will move from minority to majority within the dataset. While some algorithms are unaffected by adding new data, others most likely are. It is also to be expected that some models will perform worse during training as a result.

In a second analysis, we want to know how well the models predict low-performing children. When predicting the selective test set, we expect the oversampled models to perform at least as well as non-oversampled models. Oversampling a minority in training, however, can cause some algorithms to overfit those values when tested on unseen data, increasing the generalization error (Pykes, 2020).

Depending on the results of these two analyses, we hope to find an answer to the second part of the question, namely, which amount of synthetically generated data is appropriate to target.

Regarding question 2: When oversampling, it is crucial to implement it at the right point during training. Any wrong implementation will lead to overfitting and a false sense of security. By oversampling the training set before cross-validating, information can 'bleed' into the validation set during cross-validation (see the top section in Figure 8.5, marked in red). What does this mean? For illustration purposes, we first assume the simple oversampling strategy of duplication, where relevant observations are duplicated in order to give them more weight when training the model. Suppose the dataset is oversampled using this strategy before splitting into a training and a validation set. In that case, it is possible that identical observations will end up in both datasets, resulting in a validation set with partially already known information and thus leading to an overfitted model. More complex learning algorithms, for example, employ nearest-

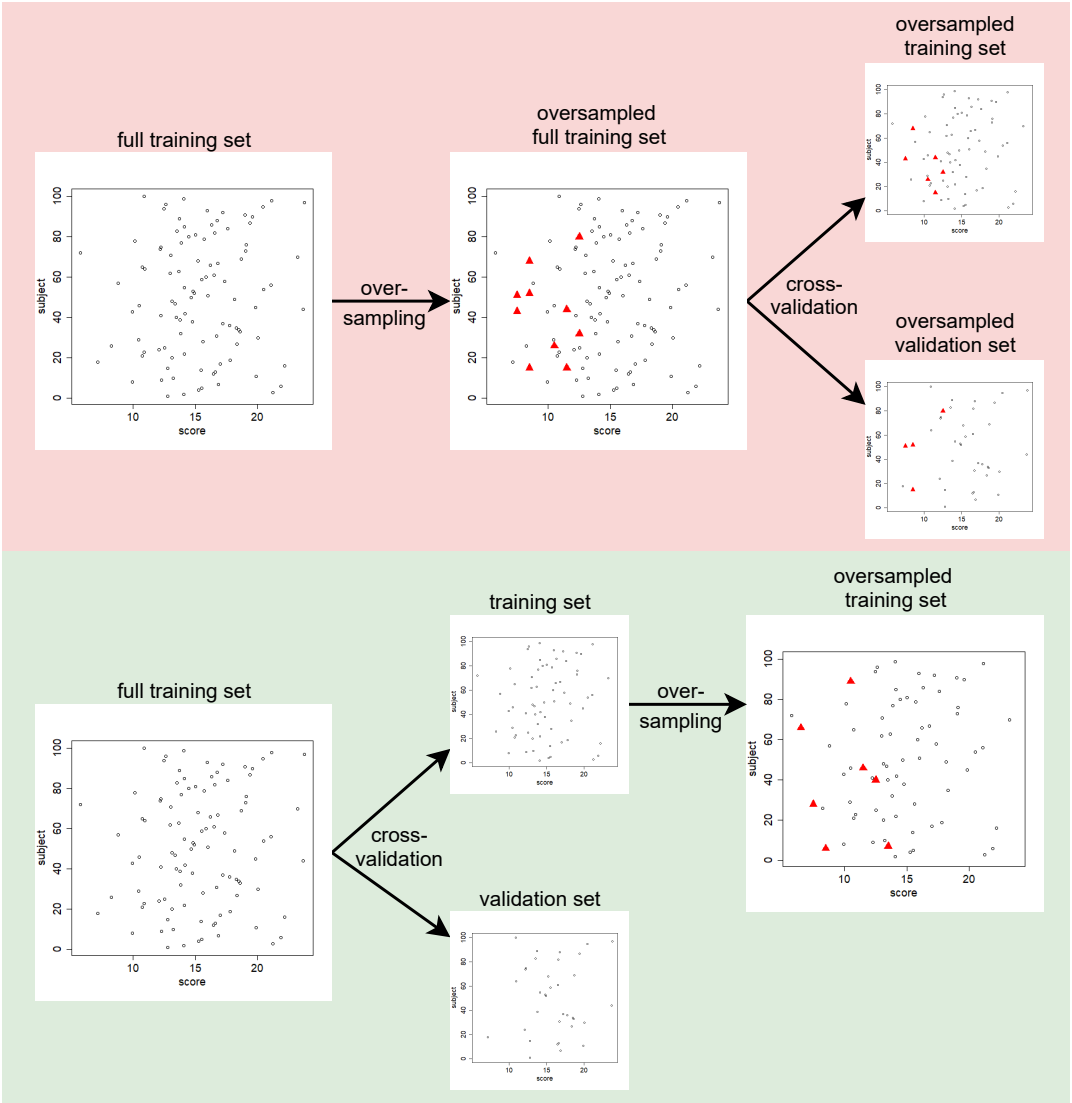


Figure 8.5: Oversampling during model training should happen only on the training set within cross-validation (lower green area) and not before (upper red area). If it happens before cross-validation, information can 'bleed' into the validation set and promote overfitting.

neighbor sampling strategies, which avoid the direct creation of duplicate observations. However, these newly created nearest neighbors still contain information from their neighbors and, if introduced in the validation set, result in the same information being present in both datasets. Since we are splitting randomly, we would expect this to happen eventually. In both cases of oversampling before cross-validation, the validation set while cross-validating may contain information from the training set and thus be incorrectly better predicted by the model (Becker, 2021).

It is vital to oversample only the training set inside the cross-validation process to avoid any bleeding into the validation set. The bottom section with a green background of fig. 8.5 shows the desired outcome of an untouched validation set and an oversampled training set to continue model training with.

Since the R-framework *caret* does not support oversampling within a cross-validated regression out of the box, we developed and implemented our own function by customizing *carets'* cross-validation function *trainControl*. The basic idea is to manually specify the indices of the train (*indexIn*) and validation (*indexOut*) sets for each fold and pass these vectors as parameters to the *trainControl*. We used a stratified subsampling method for the training and validation set. While generating these two datasets, we were also able to oversample only the train set, leaving the validation set untouched for correct oversampling inside cross-validation. A detailed workflow of this custom cross-validation process can be seen as an example in Figure 8.6.

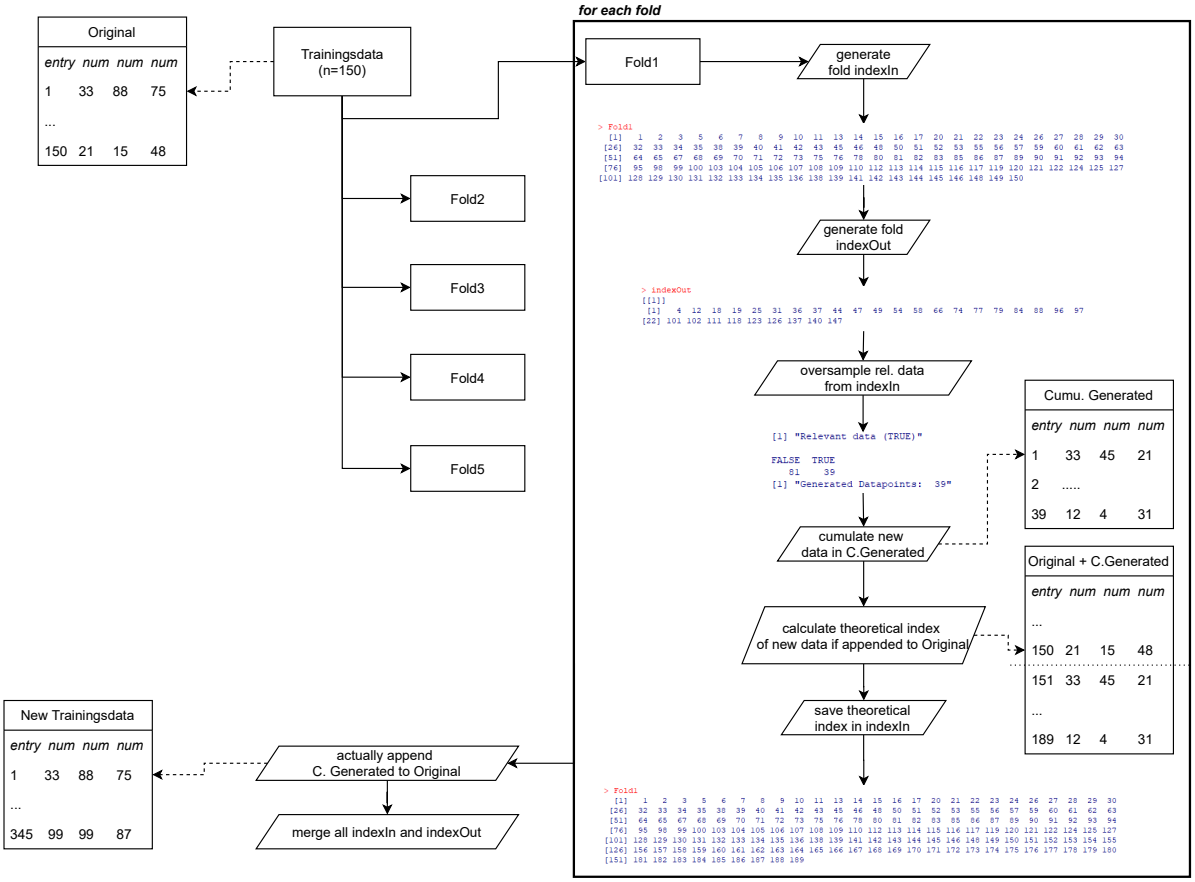


Figure 8.6: Detailed workflow of a custom cross-validation process using *carets'* function *trainControl*. This function allows for manually integrating the train (indexIn) and validation (indexOut) sets for each fold. Using this property, we integrated oversampling inside the cross-validation process correctly.

Chapter 9

Methods

Machine learning can help us understand and manipulate complex datasets. With a better understanding of the given data structure and a way to reduce or manage this complexity, we hope to develop models that utilize the data properties and maximize the prediction quality with the information available. As we have argued in favor of regression as the preferred approach before, we also have

- discussed how we could handle imbalances in our dataset,
- evaluated and decided for a cross-validation technique,
- implemented a custom two-variable stratified, single-hold-out, random subsampling algorithm as preparation for model evaluation,
- explored and integrated oversampling inside cross-validation for model training.

In this chapter, we further incorporate the aforementioned machine learning methods that help us build predictive models for reading and writing predictions, allow us to evaluate those models and the presented literacy screening overall, and propose a method to make the regression results more meaningful by including prediction intervals.

We used the statistic software R 4.2.0 (R Core Team, 2020) for all of our data processing, statistical analysis and plotting. Using the R package *caret*(Kuhn, 2008), we were able to access to implementations of relevant learning algorithms and data manipulation techniques. Other packages such as *UBL*(Branco et al., 2017), *randomForest*(Liaw and Wiener, 2002), and *forestError*(Lu and Hardin, 2021) were also essential in the implementation, evaluation and exploration of machine learning models.

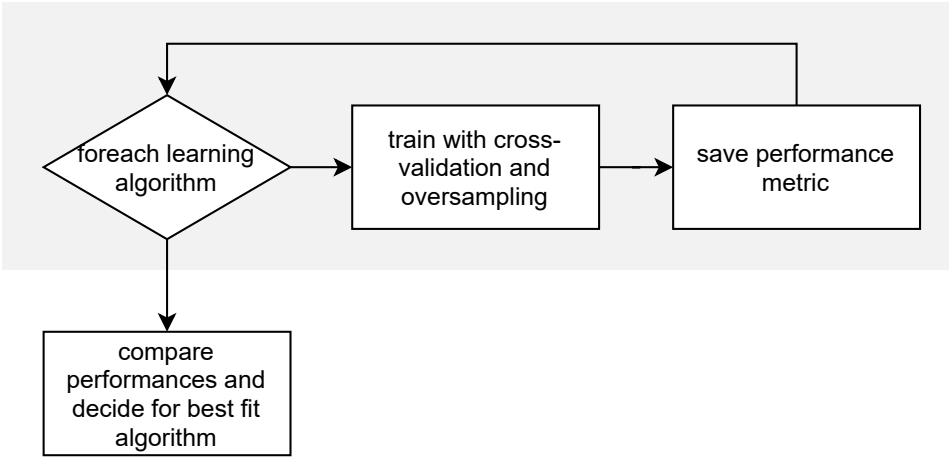


Figure 9.1: Basic process to find the best fitting model. First, train all models and then use their training performance metrics for comparison with each other.

9.1 Model training

In this section, we discuss the generation of models. Our goal is to find methods and procedures to generate separate prediction models for reading and spelling performances. In doing so, we explore how to optimize the screening data in preparation, which steps are necessary to find the most fitting prediction models, especially in the literacy screening context, and how to create and utilize a norming basis for future predictions. Our main contribution lies in adapting and applying machine learning methods to train a screening model for children with weak reading and spelling skills.

9.1.1 Finding the best model for each target variable

To find the best learning algorithm model for a given dataset, a set of suitable models is defined, trained, and compared based on their training performance. Figure 9.1 illustrates a basic blueprint for this process.

Over the course of the following paragraphs, we expand this process into a more complex structure that is guided by the properties of the screening dataset and incorporates all of the necessary components for final model training for each target variable.

9.1.1.1 Model selection

The selection of regression learning algorithms largely depends on the structure of the dataset and partially on the technical capabilities. Our model selection was primarily based on regression models that have been used in similar research settings, that are widely used for regression problems and documented in the machine learning environment, and that are optimally supported in R. With the research of Rauschenberger et al. (2020b), we found a project with similar goals that also employs machine learning technologies to predict children's future reading and writing development. Although the researchers approached the prediction as a classification, most of the models can still be applied for regression analysis. Further models could be identified, among others, via the *scikit-learn* framework, which is specifically aligned with the programming language Python but contains good explanations of the common learning algorithm models.

A list of all machine learning regression models considered for comparison and a definition of their hyperparameters used for tuning are listed in the appendix in Table C.1.

Baseline model DummyRegressor: When explaining a given dataset, it is always a good idea to start with the most simplest model (Ameisen, 2018). In the context of regression, such a model - called baseline model - would apply a very basic strategy of always predicting a constant value (e.g., mean or median) as the outcome. If a more complex model does not generalize as well to the data as the baseline, then the problem may be too difficult or the code may be flawed (Ameisen, 2018). Therefore, it is common practice to set up and train a baseline model for later comparison to other models.

Based on the *DummyEstimators* section in the *scikit-learn* framework for python (Pedregosa et al., 2011), we developed a *DummyRegressor* in R with a mean prediction strategy. Given a test and training set, the DummyRegressor calculates the mean value of the target variable in the training set and predicts this value for each observation in the test set. The predictors are irrelevant for this type of model.

Multiple Linear Regression: One of the simplest forms of regression is linear regression, which assumes a linear relationship between input and output. Since we have chosen our features carefully beforehand, we want to include all selected features and consider all their interactions and combinations of interactions. Since the features for both reading and writing target variables differ, the corresponding term for each predic-

tion turns out differently but follows the general expression:

$$output \sim f1 * f2 * f3...$$

While this consideration of all interactions covers all possible predictor combinations, it could lead to the term becoming very extensive and the model too complex to compute.

Recursive Partitioning and Regression Trees (rPart): A common technique to explore and understand the structure of a dataset is recursive partitioning. In this process, the dataset is repeatedly divided into several subsets until certain criteria are met (Zhang, 2016). The result is a decision tree that grows from the top (root) and at each node, the algorithm decides on the best-split cutoff that leads to the greatest purity in each subpartition. This procedure also reduces the effects of noisy predictors because they are most likely not selected as cutoffs (Hastie et al., 2009). The complexity parameter (cp) can be set to change the penalty for too many divisions and specify how much relative error improvement is desired from the division at the node. A higher cp value produces a smaller tree with fewer nodes, leading to coarser predictions with less overfitting. In contrast, a lower cp leads to a bigger tree with more splits and a higher chance of overfitting.
In contrast to multiple linear regression, decision trees always assume that all terms interact with each other. Every variable in the tree is compelled to interact with every variable higher up in the tree. If there are variables with no or weak interactions, this would be very inefficient. Since we only include a handful of features that are carefully selected and evaluated, we do not expect this to have a negative impact on our modeling process.
Advantages of decision trees according to Maimon and Rokach (2014) are as follows:

- Easy to read and interpret
- Robust to outliers, since they most likely will get isolated in a smaller subset, and missing values
- In principle, no specific data pre-processing is required

Random Forest: Instead of using only one decision tree to define a model, the Random Forest algorithm creates multiple decision trees that are used as an ensemble. Ensemble methods entail using many learners to improve the performance of any of them

individually. These techniques can be defined as techniques that use a group of weak learners (those who achieve only marginally better results than a base model on average) to create a stronger, aggregated learner. When combined with a randomly bootstrapped dataset for each tree and a random subsample of features to consider as cutoff at each node, we don't obtain the same tree every time and therefore reduce drawbacks such as overfitting and high variance of a single decision tree while increasing the overall performance (Breiman, 2001). With the tunable parameter *mtry* to control the random subsample of features at each node, we ensure that the ensemble model does not rely too heavily on any individual features and makes fair use of all potentially predictive features.

Gradient Boosting Machines (GBM): Boosting Machines rely on decision trees, with each new tree being fit on a modified version of the original dataset. At first, a simple decision tree is defined with equal weights for all observations. After evaluating the first tree, weights of difficult-to-predict observations are increased and weights of easy-to-predict observations are decreased. A second tree is then built upon the new weighted data. The prediction error from this new 2-tree ensemble model is then computed, and a third tree is grown to forecast the revised residuals. Unlike Random Forest, the resulting trees are dependent on each other. This process is repeated several times and is called Boosting. The general idea is to convert weak learners into strong ones.
The popular algorithm Gradient Boosting Machines builds on this idea by identifying the weak learners using gradients in the loss function instead of weights. This enables sharper calculations and tailored, optimized loss functions that can be better applied to real-world problems (Friedman, 2001).

An extension of gradient boosting is the approach of Extreme Gradient Boosting, or XGBoost (Chen and Guestrin, 2016). This modification uses a more regularized model formalization to control for overfitting, which generally improves performance. This approach, however, requires more parameters to be tuned for optimal configuration, which leads to a much higher computational expense when training the model to find the best hyperparameters.

Ridge and Lasso Regression (GLMNet): Ridge and Lasso regression are two basic strategies for reducing model complexity and preventing overfitting that may occur with simple linear regression. In ridge regression, the cost function is altered to regularize

the coefficients and decrease their impact. This allows for a reduced model complexity and less multi-collinearity. Lasso[18] regression also introduces regularization, but this can result in selected coefficients being reduced to zero and therefore are neglected for the evaluation of the output. Generally speaking, this can reduce overfitting and help us in feature selection(Bhattacharyya, 2018). Integrating both algorithms into one model and introducing the *alpha* parameter to define the mixture of both methods allows for a model that can find and incorporate the best combination of both approaches.

9.1.1.2 Metrics for model comparison

When looking at metrics to compare different regression models, we first need to define at what stage of the modeling process we want to compare and differentiate between performance and evaluation metrics.

Performance metrics can be derived by assessing how well a model predicts the validation set using the training set during training. Subsampling and cross-validation methods at this stage can help to generalize the performance results and make them more reliable for comparison. This approach is considered good practice, as it enables formal comparisons of models prior to the involvement of the test set (Kuhn and Johnson, 2019). Using the test set to calculate evaluation metrics for the purpose of model selection violates the intended use of this dataset. It should remain unseen by all models during the selection and tuning process, ensuring that once a model is finalized, it can be objectively evaluated on this untouched data to accurately gauge its performance on new information.

The most common metrics to consider when comparing regression model performances are root mean squared error (RMSE), residual standard error (RSE), mean absolute error (MAE), r squared (R2) - respectively, the more unbiased adjusted r squared (AdjR2) - Akaike's InformationCriteria (AIC) and Bayesian Information Criteria (BIC) (Kassambara, 2018). However, when using cross-validation as we do in our screening, the RMSE is recommended (Kassambara, 2018) because it is more sensitive to the occasional large error when compared to other error measurements (Nau, 2021) and allows for better interpretation of the error rate since it is measured in the original unit. We considered these elements to be relevant to our decision-making and therefore chose the RMSE as our only performance metric to compare between models.

[18]lasso=least absolute shrinkage and selection operator

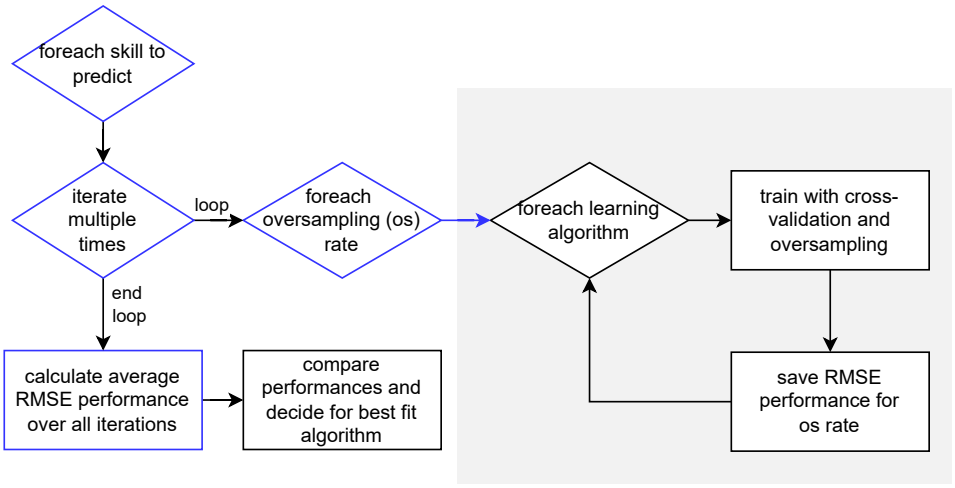


Figure 9.2: Extension (in blue) of Figure 9.1 with the additional calculation of the average RMSE and different oversampling rates.

9.1.1.3 Detailed model training for each target variable

To improve and consolidate the results of the model training, we extend the process presented in Figure 9.1 based in part on the findings from the previous sections:

- Run a separate model training for each target variable.
- Average the RMSE over multiple repetitions of model training with a different training set each time to gain more robust performance results.
- Test different oversampling percentages and compare performances to find an appropriate ratio (see Section 8.3.4).

The resulting model training process is displayed in Figure 9.2. It is worth mentioning that the oversampling of observations from low-performing children is integrated into the k-fold cross-validation process during model training. This custom algorithm is described in Section 8.3.4.

Although the complexity of the model training and therefore its computational effort is drastically increased with the integration of the aforementioned aspects, this procedure fulfills many tasks at once: Compute and analyze the effect of different oversampling rates and produce stable training performances for better model comparison.

A closer look at our final model training highlights a few more methods, which are briefly listed below. They help us to prepare the data correctly and to make our analysis reproducible. A visual representation of the extended model training in detail can be

seen in Section 9.1.1.3.

Prepare data: As evaluated in Part B, we found that each target variable has slightly different predictors. According to these findings and depending on the current target variable to be processed, we prepare the data with the necessary independent variables.

Calculate oversampling threshold: Derived from the fact that we want to identify children with weak reading or spelling performances, only the values in the range of low-performing children should be oversampled. This requires a threshold that decides which values to oversample and which not to. An alternative approach - the relevance function according to Branco et al. (2017) - could not be successfully implemented using the UBL framework. The SLRT II and SLS 2-9 raw values at one standard deviation below the mean were eventually taken as thresholds. For the threshold calculation, only the data of children who had indicated German as their native language were used. With this approach, we follow other research to identify low-performing children in an educational context, such as Mayringer and Wimmer (2014)[19].

Set random seed: A flexible yet comprehensible random seed generation is necessary to increase the reproducibility of the process and still produce different data each iteration, for example, when splitting training and test sets. This system is needed when (1) calculating the subsamples for the learning set, (2) training with cross-validation and (3) oversampling within cross-validation.

Five iterations: Since we already implemented multiple iterations for the model training in regards to different oversampling rates and different random seeds for data splitting (further explained in Section 8.3.4 and Section 9.1.1.3), we decided against the inclusion of multiple repetitions for the k-fold cross-validation, which would further increase computational complexity (see Section 8.3.3.1).

9.1.1.4 Decide for the best performing model

After all models are trained at each oversampling rate, we compare the models throughout all oversampling rates using the corresponding average RMSE training performance metric. We can then determine which model works best for each target variable and

[19] A more detailed explanation to the one-standard-deviation-threshold can be found in Section 2.1.1.

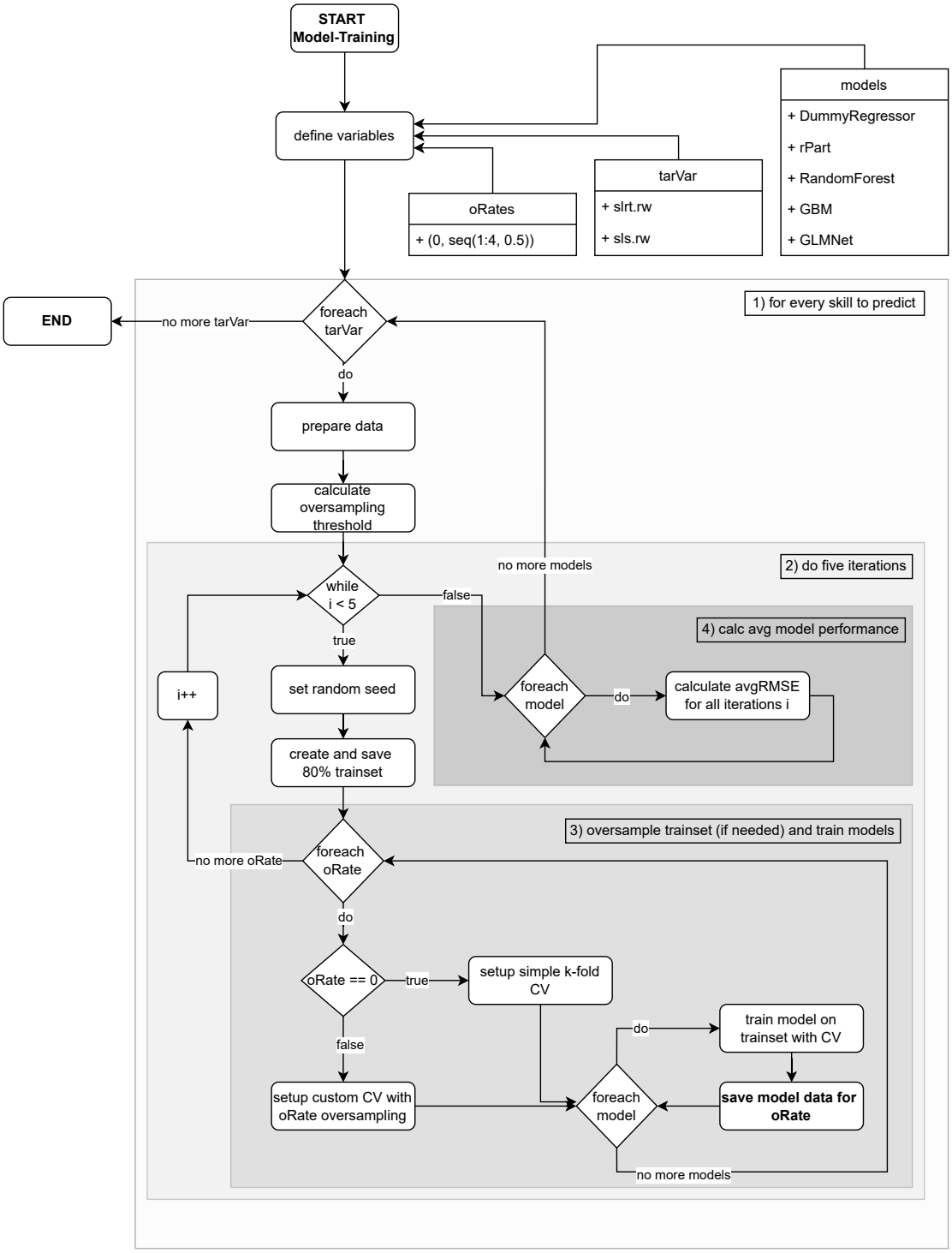


Figure 9.3: Flowchart of the model training process. As part of the workflow, multiple models are trained for each of the two target variables and cross-validated on multiple training sets with different oversampling rates. The model performance is averaged over five iterations to get reliable data.

at what oversampling rate it does so. The goal at the end of the decision process is to decide on a learning algorithm, the model of which can then be used for evaluation.

9.1.1.5 Decide for the best oversampling rate

There is no fixed formula for determining an oversampling rate. A too-high oversampling rate should be avoided because it can lead to overfitting and reduce the model's generalizability. It is therefore advisable to test and compare different oversampling rates, starting e.g., with a rate of 100% up until 400%, and observe how the model's performance changes. The optimal oversampling rate may vary depending on the specific dataset and the problem you are trying to solve.

9.1.2 Generating a standardization basis from prediction

Pre-tests have shown that a regression to the mean can be assumed for the model predictions. This means that, for example, the predicted reading value of a child cannot be directly compared with the original reading value since the model tends to estimate the values towards the mean value in order to minimize the prediction error. The values predicted from the models are not to be compared with the raw values from the manuals of the respective SLRT or the SLS tests. Figure 9.4 displays an example from pretest data, in which the measured target variable of children is compared to their predicted values by a Random Forest model. Although the distributions are different, a child whose actual reading performance is in the top 10% compared to all other children should also be ranked within the top 10% in the new distribution of the predicted values. If, in a future scenario, models from this thesis should be applied to literacy skill predictions, the results of the predictions have to be assigned relatively to the new standardized basis instead of absolutely to the original distribution.

In this section, we address the methods required to create a new distribution of the target variables from the model predictions in order to calculate a new threshold to distinguish low-performing children from all others for future evaluations.

9.1.2.1 Predicting performances of children

With our best-performing models at hand, we predict the target variables for all children. Since our goal in this step is to create a large and diverse output of our models, overfitting is irrelevant and predictions can be made even on the children used for training. In the

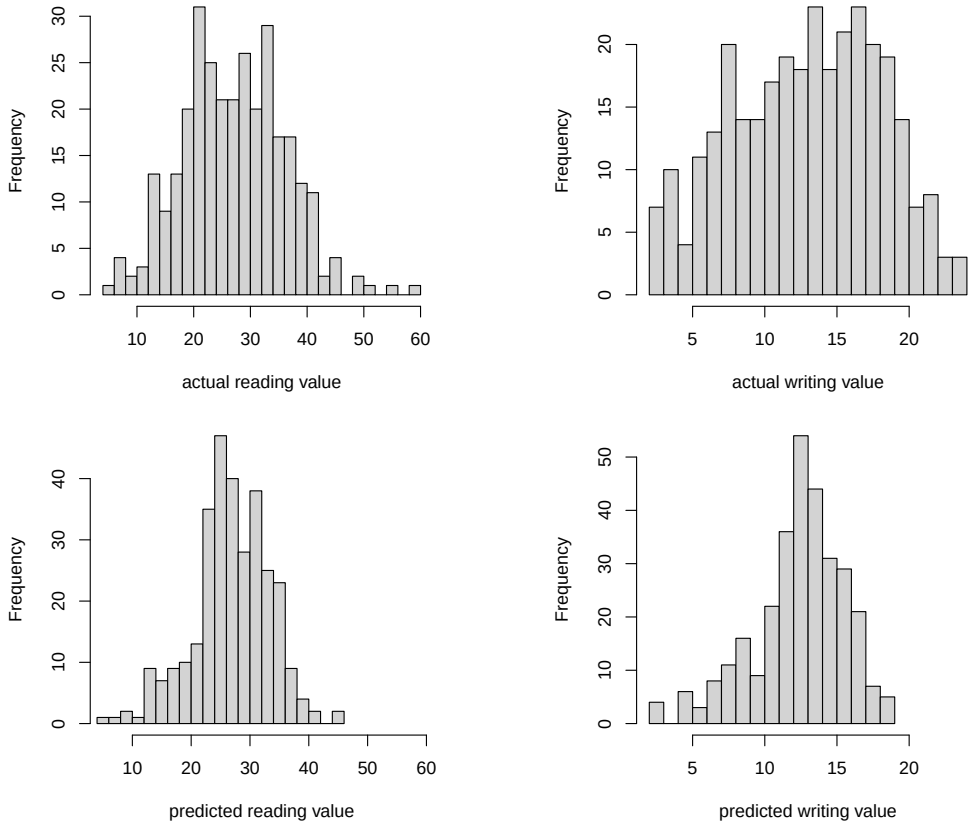


Figure 9.4: The histograms illustrate the regression to the mean effect for both predicted reading (left) and spelling (right) values by a Random Forest model. The data basis is a preliminary analysis and does not correspond to the final data set (n=306).

context of this work, the children's predictions are adopted as the normalization basis for future predictions with these models.

9.1.2.2 Calculating a new threshold for low-performing children

The threshold used to determine low-performing children in training (e.g., to define the oversampling range) is defined by one standard deviation below the mean of the measured target variable values of German-speaking children (further described in Section 2.1.1). Since the predicted values, as described in the previous section, are being regressed towards the mean and therefore differentiate from their original distribution, a new threshold for a later evaluation and possibly classification of the children between weak readers or weak spellers has to be calculated. The predicted values of only German-speaking children are again used as the data basis for this calculation.

9.2 Evaluation

Following the model training, we are left with two separate models: one for predicting the reading and one for predicting the writing score. This section addresses the methods needed to evaluate our screening as a whole. So far, we have mainly focused on the separate creation and interpretation of the two types of models. That is, until the joint evaluation becomes necessary and we want to compare our version of a dyslexia screening with other screenings. Since the used R-framework *caret* does not provide many methods for such a special, two-target-variable case, manual adjustments and new implementations are necessary. In the next sections, we explore the path to a unified evaluation of our screening by evaluating the individual models and classifying the results to eventually create unified metrics for comparison with other screenings.

9.2.1 Procedure to evaluate the screening

The evaluation of the screening can be split up into two steps once a fitting model is found and trained:

1. Evaluate the best-performing model for each target variable with the use of sub-sampling methods
2. Evaluating the screening using post-classification
    - Classify results, create confusion matrix and classification metrics

- Compare results with other screenings

In the following sections, we will go into more detail about these steps and the methods required to do so.

9.2.2 Evaluating the best-performing model for each literacy skill

After deriving an oversampling rate and the best-performing model for each literacy skill from model training (see Section 8.3.4), these models must be evaluated to analyze their validity. In evaluation, the test set, which is a subsample from the learning set, is presented to the already trained models as new and previously unseen data. The basic idea is to make predictions with the model at hand based on the unseen test set and measure how accurate they are when comparing them with the actual data.

We need to separate the evaluation process from the training process described in Section 9.1. Model training is done to generate a model for comparison and analyze different oversampling rates. The general workflow of model evaluation is similar to that of our model training with a few, yet important, differences.

- The step where the learning set is split into train and test set is outsourced before the model training. This means the split does not occur when training the model.
- Since we have already decided on an oversampling rate, the complexity of determining one is eliminated. We train both target models accordingly by oversampling the children with literacy performances smaller than 1SD.

For evaluating the model, we need to make sure to predict both outcomes for each child in the test dataset. If we create the test and training data only when training the models, and thus separately for both target variables, there will be children with only one prediction for reading or writing. No overall statement can then be made about the overall predictive quality of the screening. This ultimately requires the custom subsampling method described earlier in Section 8.3.2.2 that is stratified based on two variables to generate the test and training set before the model is trained. We decided on a split ratio of 70% training and 30% test set so that we have a relatively large test set for optimal evaluation of the models.

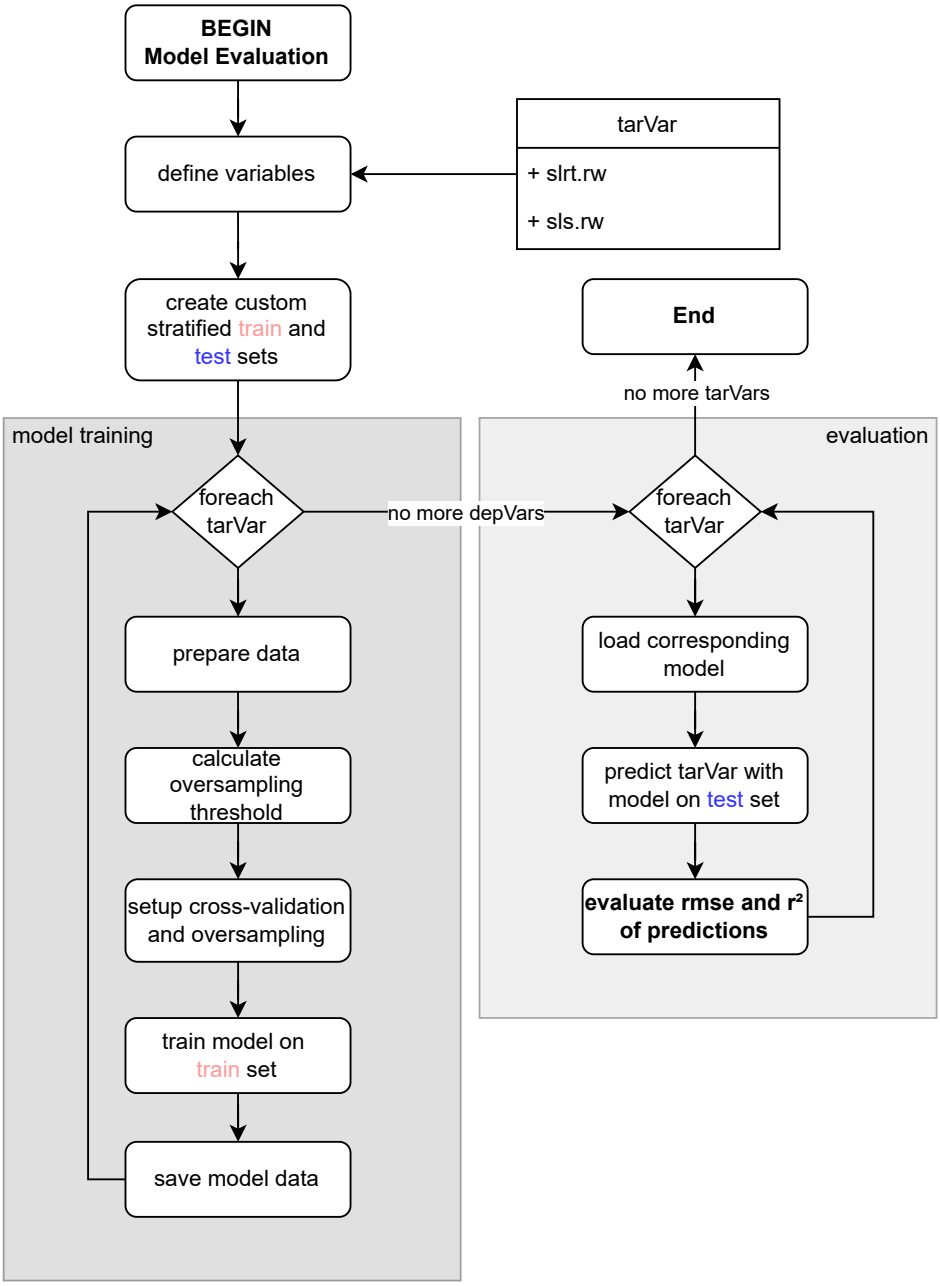


Figure 9.5: After splitting the data into a train set and a test set using our custom stratified subsampling method, for each literacy skill, we prepare the data, calculate the threshold on the basis of all available data to apply oversampling on and finally train the model using k-fold cross-validation. The model is saved and then used to predict the reading or writing outcome of the children in the test set. The resulting RMSE and r^2 metrics are calculated and stored.

Table 9.1: Confusion matrix for a literacy screening.

		Actual state	
		affected	not affected
Prediction state	at-risk	True positive (a)	False positive (b)
	not at-risk	False negative (c)	True negative (d)

9.2.3 Evaluating the screening using post-classification

This subsection briefly describes the methods needed to make the screening comparable to other screenings. To do this, we must first transform the screening task into a classification problem and define classes, as well as metrics for that. The results obtained from the evaluation of the regression models can then be classified. We are guided by the meta-study of Marx and Lenhard (2011) and the metrics defined therein to assess the different screenings better.

9.2.3.1 Metrics for screening comparison

In the meta-study by Marx and Lenhard (2011), a selection of screenings for German-speaking children were analyzed and re-evaluated with universal metrics. The authors first defined and calculated a set of indicators to measure the predictive performance and then used these metrics for comparison. The calculations are based on the screenings' classification results and the resulting confusion matrix. An overview of the measurement units used in the meta-study with their description and calculation can be seen in Table 9.2. Derived from the confusion matrix, sensitivity and the positive predictor value are to be considered as key performance indicators. However, other common metrics such as specificity and hit rate[20] should be treated with caution when the dataset is unbalanced. For dyslexia screenings, the specificity is usually relatively high, as significantly more children do not have dyslexia. Similarly, a high hit rate suggests good prognostic validity but must always be considered in relation to the random hit rate. This relation is represented by the Ratz index (Marx, 1992), which initials translate from German as the relative increase in hit rate compared to the random hit rate (German: Relativer Anstieg der Trefferquote gegenüber der Zufallstrefferquote). Simply put, it indicates how much the procedure beats chance.

[20] In the context of machine learning, also known as *accuracy*

Table 9.2: Metrics used in meta-study by Marx and Lenhard (2011), their calculations and what they measure in the context of a dyslexia screening. The variables for the calculations are based on the confusion matrix template as shown in Table 9.1. Such confusion matrices are the result of a prediction classification algorithm.

Metrics	Description	Calculation
Sensitivity	Indicates the proportion of actual children with problems that have been identified as at-risk children through screening.	$a/(a+c)$
Specificity	Indicates the proportion of actual children without problems that have been identified as not at-risk children through screening.	$d/(b+d)$
Positive Predictor Value (PPV)	Indicates the probability that a child who has been classified as being at risk will actually develop problems.	$a/(a+b)$
Hit Rate (TQ)	Indicates the proportion of correct classifications by the screening, i.e., the proportion of correct positive predictions and correct negative predictions in all cases.	$(a+b)/(a+b+c+d)$
Max. Hit Rate (max. TQ)	Maximum possible hit rate. This only differs from the Hit Rate if the amount of actual diagnosed children with dyslexia differs from the expected amount.	$1-\|(b-c)\|/(a+b+c+d)$
Random Hit Rate (ZTQ)	If we take the frequency of occurrence of dyslexia and randomly draw this proportion of individuals, there is some probability of correctly identifying individual affected individuals.	$[(a+b)*(a+c)]/(a+b+c+d)^2+$ $[(c+d)*(b+d)]/(a+b+c+d)^2$
RATZ Index (Marx, 1992)	This value indicates the extent to which the prediction generated by the screening exceeds a random prediction.	$(TQ-ZTQ)/(max.\ TQ-ZTQ)$

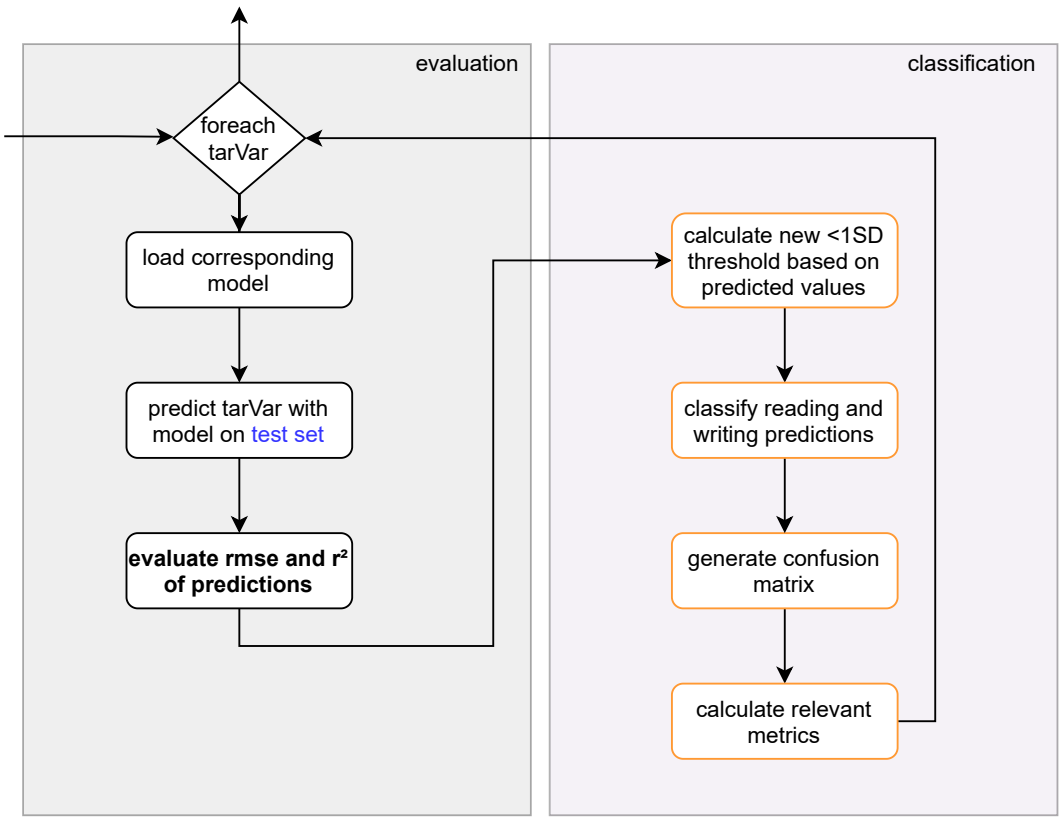


Figure 9.6: Extension of the workflow described in Section 9.2.2 for the evaluation of individual models in order to be able to evaluate the screening as a whole. For this purpose, regression model results are first classified and then converted into comparable metrics.

In order to use the same metrics as Marx and Lenhard (2011) to evaluate our literacy screening and compare it with other screenings, we need to classify our results from both target variable regressions.

9.2.3.2 Classifying results

During the evaluation process, the model's quality gets measured by making a prediction of the target variable on unseen data and comparing its result to the actual target variable of the data. To further classify these predictions, we first define the classes into which we wish to divide them. Since we want to identify children that are *at-risk* of developing reading and writing weakness, our class labels are *weak* and *not weak*. The workflow adapted to the classification can be seen in Figure 9.6.

Children are classified as a *weak* reading and spelling candidate if their performance is

below the respective threshold. To classify children based on their reading and writing scores, measured by the SLS and SLRT tests during the field study, we defined the threshold as one standard deviation below the mean of the target variable ($< 1SD$), considering only German-speaking children (more details in Section 2.1.1). The same applies to the classification of the predicted values and all future predictions by the models. Due to the regression to the mean effect , the new threshold described in section Section 9.1.2.2 must be used for these values to classify them into the defined classes.

Once classes and thresholds are defined, predictions made on the test data can be classified and converted into a unison class for both target variables. Eventually, they can be used to create a confusion matrix. With this data at hand, we can calculate all aforementioned evaluation metrics and compare the screening with other screenings.

9.3 Interval estimation

An essential part of a literacy screening is the generation of meaningful output for the target audience. Many systems create a binary classification, answering the question: is my child *at-risk* or is *not at-risk* of having dyslexia, meaning having a weak literacy score? In chapter Section 8.1, we discussed that we can reduce the loss of information that can result from strict classification by using regression. By using the data presented in this paper as a normalization basis for future predictions, we can make even more accurate regression predictions and realize the full potential of the information available to us. In this chapter, we would like to present a possibility, which, to our knowledge, is unexplored in the literacy screening context, to increase the informative power of regression results by including uncertainty in the form of prediction intervals.

9.3.1 Prediction intervals and their benefits in a literacy screening

Similar to the more well-known confidence interval, prediction intervals allow the estimation of a range of values in which the true population parameter will be present with reasonable certainty. Confidence intervals indicate how accurately a parameter of interest, such as a mean or regression coefficient, has been calculated. Prediction interval, on the other hand, indicates where you may anticipate the predicted value to be. Because

the prediction interval must account for both uncertainty in predicting the population mean and random variation in individual values, it is always wider than a confidence interval (Olive, 2007). By integrating the prediction interval, we think the power and interpretation of the screening regression results can be greatly improved.

For example, with a fully trained reading model and the dataset of a previously unknown child, we can predict not only the probable raw reading score but also a range from the upper to the lower bound where the target score will be present with a set probability of, for instance, 95%. Parents, teachers, therapists, or others interested in the predictions are then able to estimate the severity of the situation better and make a more informed decision for further action (e.g., performing further diagnosis or starting therapy).

9.3.2 Implementing prediction intervals

The prerequisite for using prediction intervals are the fully trained models and the oversampled train set used for training. We decided to use the *quantForestError* function for Random Forest models from the R-package *forestError*(Lu and Hardin, 2021) to calculate the prediction intervals. Given our instance of a final reading or writing model, our training set, and an appropriate *alpha* value for the desired type-I error rates, the raw score prediction and its prediction intervals can be computed for each new child. For illustration, before training the final models, we extracted three children whose performance we would like to predict from the entire data set. We set the *alpha* value of the *quantForestError* function to 0.2, resulting in an interval of 80%. In this way, our prediction interval is acceptably narrow. Together with the final model of the respective target variable and the train set used for this model, we generate the mean prediction value, the upper 10% and the lower 10% prediction bound.

9.4 Contributions

In this chapter, we first defined the workflow for our model training and then described the procedure for evaluating these models and the entire screening since we want to compare it to other screenings. In an effort to provide a more focused and, in our opinion, better interpretation of the regression results of the reading and writing models, we also propose to include uncertainty by adding the prediction interval to each prediction made. Our contribution in this regard can be summarized as follows:

- Comparing different learning algorithms based on our screening dataset
- Analyzing different oversampling rates for the optimal amount of additional synthetic data
- Generating a standardization basis for future predictions according to our data and our models
- Classification of regression results on the basis of a specially defined threshold to eventually generate comparable metrics
- Proposal for an approach that could make the regression results more meaningful with the use of prediction intervals

To evaluate our screening and make it comparable to the other conventional screenings, we need to convert the results from our regression models into classification results. For this, we developed the following methods:

1. Custom stratified subsampling method dependent on two target variables (see Section 8.3.2.2)
    - Having a stratified distribution of two target variables allows us to use the same test set and sample of children for model evaluation and, eventually, screening evaluation.
2. Classifying results of two separately evaluated prediction models
    - Given a new threshold to separate children into weak and non-weak literacy skills, this classification allows for the screening to be compared and evaluated as a whole.

Chapter 10

Results

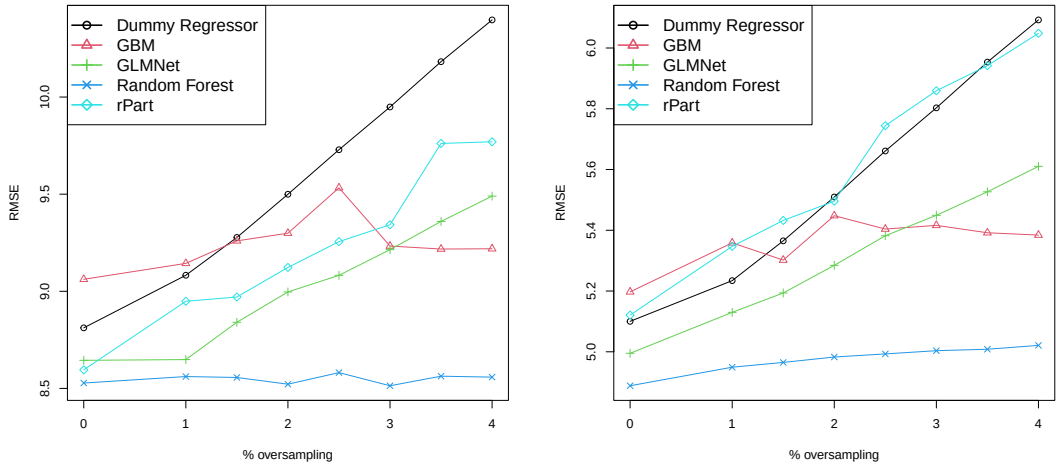
10.1 Model training

For the final model training, we set up and compared five machine learning models: a dummy regressor model, a recursive Partitioning and Regression Tree model (rPart), a Random Forest ensemble, a Gradient Boosting Machine model (GBM), and a ridge and lasso regression model (GLMNet). We could not calculate the multiple linear regression model for either of the two literacy skills. This is due to the high number of predictors and the possible interactions between these predictors, leading to the model being too complex and thus too computationally intensive.

As data basis for the training, we only accepted children with a complete data set, consisting of all screening tasks, the reading and spelling tests, and the covariates, i.e., CST and age. Using this data set ($n = 272$), we calculated the 80% stratified training set ($n = 219$) and trained all models on it. At each iteration, we used only 80% of the data as training set, so we could create a new composition of the data for the next iteration with a different random seed.

10.1.1 Comparing model performances

Figure 10.1 shows the training performance of the models starting from a training set with no oversampling to a training set with an oversampling rate of 400% in total with a 50% rate interval at each step. The graph describes the average performance of the models across all five iterations of the training.



(a) Training performance of all models for predicting reading skill.

(b) Training performance of all models for predicting spelling skill.

Figure 10.1: Overview of model performance for both literacy skills when training on data without oversampling up to 400% oversampling rate, starting at 100% in 50% increments. The performance of the model is worse the higher the RMSE value.

The Random Forest model performed best for both literacy skills and showed the lowest RMSE on average across all oversampling rates and iterations. The data basis for Figure 10.1 can be found in Table C.2.

For both literacy skills, we saw that the RMSE of the dummy regressor, the GLMNet model and the rPart model steadily increased with higher oversampling rates. The GBM and Random Forest models, on the other hand, showed a roughly constant performance between 0% and 400% oversampling for the prediction of reading skills. For the prediction of spelling skills, the RMSE of these models increased to a very small extent, but still far less compared to the other models.

For each type of model, we have set up and trained 80[21] models. To find the best set of hyperparameters, we had to perform individual calculations for each of these models (see Table C.1). However, a list of the hyperparameters used in the training for each model is beyond the scope of this thesis. A list of the hyperparameters of the final Random

[21] For both literacy skills, we iterated five times through all oversampling rates (0%, 100%, 150%, 200%, 250%, 300%, 350%, 400%).

Forest model can be found in its evaluation in Section 10.2.

10.1.1.1 Model training

As expected, the dummy regressor - our baseline model - showed worse RMSE values with an increasing oversampling rate. Adding data points in the lower literacy performance spectrum and the associated skewed distribution increased the error rate of the model's prediction to the mean value. The rPart regression model performed about as well as the baseline model in predicting spelling and performed similarly but slightly better on average in predicting reading skills. Therefore, the rPart model does not seem to be suitable for the present use case and the data basis. The GLMNet model showed a comparable trend - increasing error rates with increasing oversampling rate - but with a better RMSE performance on average, also making it not a real contender for future predictions.

Although the performance of the GBM model remained relatively constant across all oversampling rates compared to most other models, the error rate was above average. Thus, the model dealt quite well with the increasing number of literacy-weak children in the training set, but in general, the model is not particularly well suited for predicting literacy skills, given our predictors.

Having the lowest RMSE and the most consistent performance throughout different oversampling rates, the Random Forest model was the most suitable prediction model for predicting both literacy skills. The ensemble of decision trees seemed to handle the data complexity best, given challenges like interactions between predictors, an imbalanced data set and outliers.

10.1.1.2 Oversampling and its optimal rate

Training the Random Forest model for both literacy skills on oversampled data compared to non-oversampled data improved its performance when evaluating the trained model. In Figure 10.2, we see lower prediction errors when testing the model, that is trained on oversampled data, on a new and unseen test set, that includes only low-performing children ($< 1SD_{reading}$, $n = 19$; $< 1SD_{spelling}$, $n = 19$). We therefore can confirm that the aimed approach, to oversample only the data range of interest, positively affects the prediction of underrepresented children with reading and spelling difficulties.

Pinpointing a specific oversampling rate seems rather tricky since the performance re-

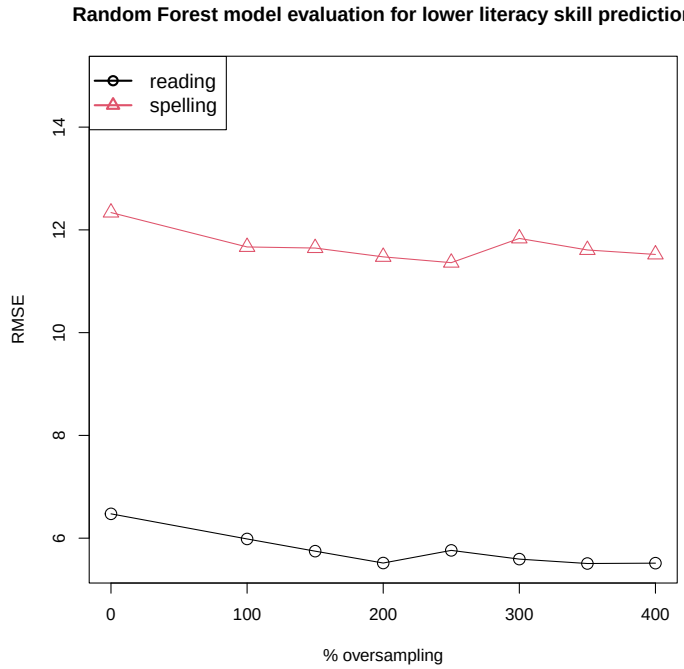


Figure 10.2: Evaluation of the Random Forest model for both literacy skills in predicting children explicitly with reading and spelling difficulties (respective raw value $< 1SD$). We trained multiple Random Forest models on data sets with and without oversampling, ranging from a rate of 0% to 400%.

sults with the Random Forest models trained on oversampled data are close to similar. To keep the rate in a reasonable range and not to skew the data too much, we decided to oversample low-performing children at a rate of 200% for the final model evaluation.

10.2 Evaluating the Random Forest models

10.2.1 Random Forest training and regression evaluation

The training data for our reading skill regression models consisted of data from 190 children collected and 312 data points oversampled in the low-performing area using an oversampling rate of 200%, giving us a total of 502 samples. Training the spelling skill regression models showed the same amount of generated samples, however, different folds and split points where used during cross-validation. For both models, the test set

Table 10.1: Evaluation results of the Random Forest Model, used on the test set.

Literacy skill	RMSE	r^2	MAE
Reading	7.92	0.18	6.56
Spelling	4.63	0.12	3.90

to evaluate consisted of 82 children (70:30 split).

The evaluation results of the regression model for both literacy skills can be seen in Table 10.1. The RMSE can be read such that, on average, the predicted value of the reading test raw value deviates by 7.92, respectively 4.63 for spelling prediction.

We used the trained model to calculate the new threshold of the predicted values. The basis for this was the complete, non-oversampled test set, and we predicted both literacy skills using the trained models. We have chosen the test set as the data basis for calculating the thresholds, as this data is free of bias, especially with regard to overfitting. With the new distribution of our predicted values at hand, we were able to calculate the new thresholds to classify children as weak readers and writers, being at 24.11 for the reading-dependent variable and 11.27 for the spelling-dependent variable. The resulting distribution can be seen in Figure 10.3.

10.2.2 Random Forest classification and screening evaluation

Given the new thresholds, we were able to classify the test data results into two classes (*weak* and *not-weak*) and furthermore evaluate the screening in its entirety, as well as compare it to other screenings.

Given a binary classification of the original target variables, we were able to calculate the confusion matrix shown in Figure 10.4. According to our model, the classification was correct for ten children who were classified as weak in reading or writing. Nine children were incorrectly classified as weak readers or writers, while seven children were incorrectly classified as non-weak in those skills. Most children, 56 in total, were correctly classified as non-weak readers and writers.

To compare the screening results with other dyslexia screenings, we calculated the metrics suggested by Marx and Lenhard (2011) in their meta-study (see Table 10.2).

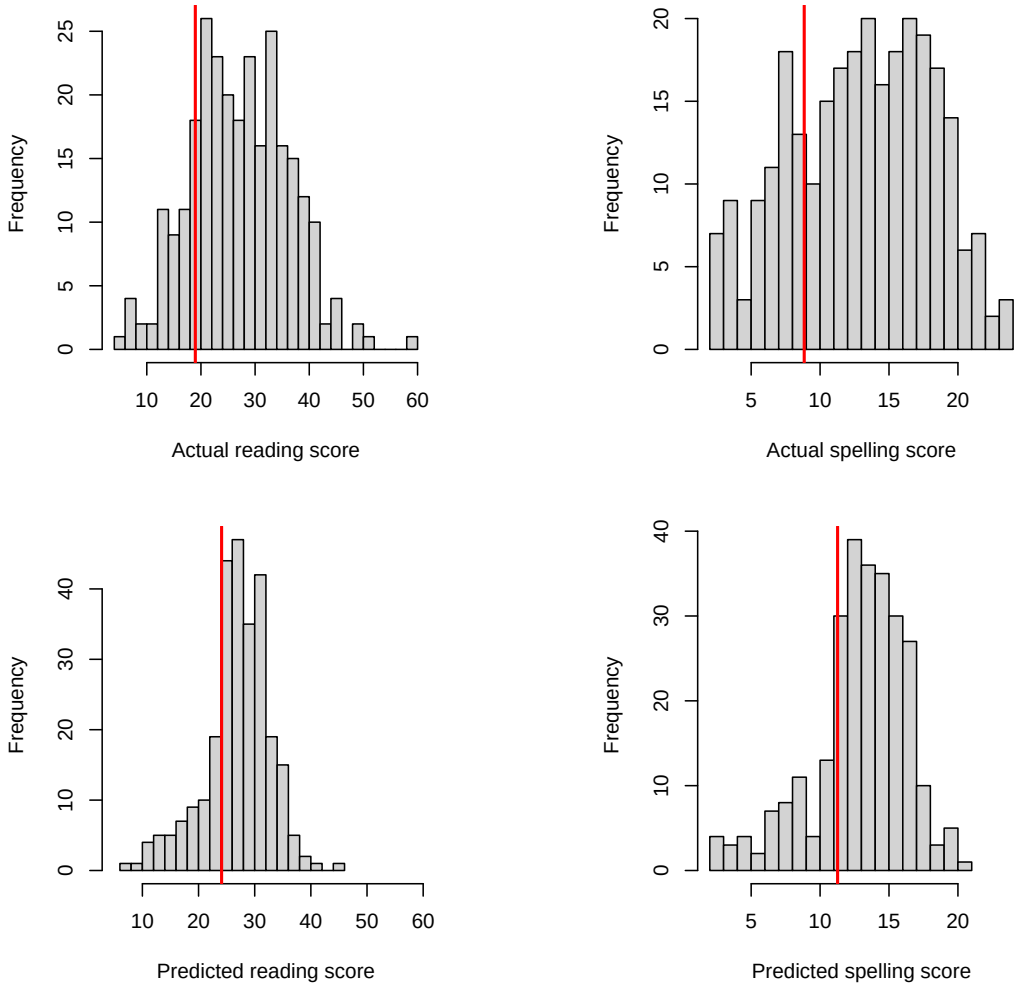


Figure 10.3: Distributions of the actual and predicted literacy skill scores in the complete dataset ($n = 272$), predicted by our Random Forest model. The effect of regression to the mean is visible in the distribution as well as in the shift of the thresholds (marked in read).

Table 10.2: Performance results of our screening according to the metrics proposed by Marx and Lenhard (2011) in their meta-study of German dyslexia screenings. PPV stands for the positive predictive value, and NPV for negative predictive value.

Sensitivity	Specificity	PPV	NPV	TQ	ZTQ	RATZ
0.59	0.86	0.53	0.14	0.80	0.66	0.46

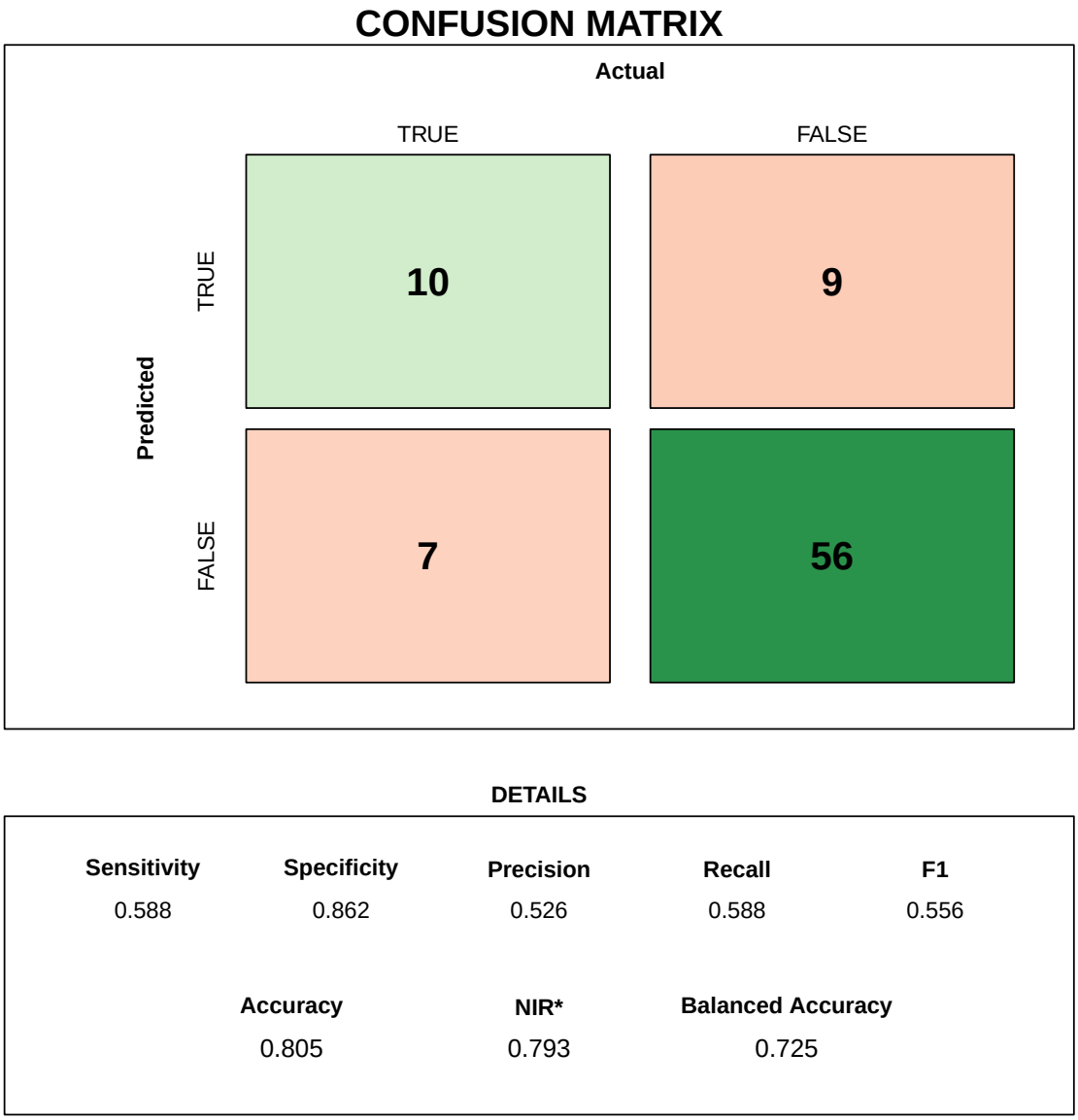


Figure 10.4: Confusion matrix of the screening when the binary classes of the predicted values are calculated using the new threshold and compared with the original classes. The basis for this evaluation was the test set, which consisted of $n = 82$ children. The results of the evaluation are listed below.

A specificity of 0.86 indicates that the model correctly identified 86% of individuals without dyslexia. However, specificity can be easier to achieve in an imbalanced dataset because there are more negative cases. The model's sensitivity is lower at 0.59, suggesting it doesn't perform as well in correctly identifying individuals with dyslexia. The random hit rate *ZTQ* of 0.66 relativizes the on-paper good hit rate *TQ* of 0.80 since a lot of cases can be explained by chance alone. As a result, the RATZ index is calculated at 0.46 and displays moderate results. Due to the small number of children in the test set and the even smaller number of literacy-weak children in the test set, the classification results shown here should be treated with caution. The data basis is too small to derive meaningful results.

10.3 Discussion

We assume that the significance of the predictors might have been better with a more meaningful data set. Among other things, the exclusion of children who showed an unusual pattern of responses to the tasks would have contributed to this. Due to a technical error in the recording, we were unable to do so for the SST.

10.3.1 Model selection

Unfortunately, training the Extreme Gradient Boosting (XGB) models was too tedious due to the high complexity and computational time involved in calculating the hyperparameters using the *caret* framework, so we refrained from doing so. Similarly, we could not compute the multiple linear regression models because the interactions between predictors were too strong.

10.3.2 Classification

We used the test set as the data basis when calculating the new thresholds to classify children based on their predicted scores. As mentioned earlier, we did this to avoid the introduction of bias in the form of overfitted values. If we had predicted the scores of all children and used these values as a data basis for threshold calculation, we would have had more data and thus possibly a better generalization available, but we also would have overfitted 70% of the predictions of the children's score. To us, the tradeoff for less data versus more uncertainty by including overfitted data seemed reasonable hence, we

decided for the test set as a data basis. Nonetheless, a larger sample of the test data would have been desirable.

Assuming the goal is to classify children's performance into the binary classes *weak* and *not weak* using the presented Random Forest models, the models should be trained on our dataset and used to predict children's outcomes of a new and unknown dataset. The values then predicted can be used as a norm basis to calculate new thresholds.

10.3.3 Oversampling

By adding oversampling to model training, we hoped for a greater improvement in predictive performance in the area of children with weak literacy skills than we saw in Figure 10.2. Although we found a consistent decrease in RMSE error rates when the oversampling rate was increased, the reduction in value was relatively small. By introducing data points in the low-performing areas, we shifted the imbalances of the learning data. As we know from Batista et al. (2004), tree-based models, especially Random Forest models, can handle imbalances inherently better than other models. Therefore, the original goal of weighting the relevant areas by oversampling to increase model prediction for those areas did not influence the Random Forests prediction power since it tried to balance out imbalances.

10.3.4 Prediction intervals

To illustrate the prediction intervals, we randomly drew three children from the dataset and used the remaining data ($n = 269$) to train a Random Forest Model for each literacy skill with an oversampling rate of 200%. With models for both literacy skills, the training data, and a fixed α of 0.2 at hand, we were able to calculate prediction intervals for each of the three held-out children. The results can be seen in Table 10.4. In it we see the predicted literacy score for each child as well as the lower and upper bound of the 80% prediction interval.

With the prediction intervals, it is no longer necessary to approach the whole procedure of literacy screening as a classification problem. In our example case for $Child_a$, the spelling Random Forest model predicted a score of 11.51, where it was confident to a degree of 80% that the value is between 10.75 and 17.41. Due to the potential for improvement in the implementation of the tasks and the general screening structure, the predicted values are still quite inaccurate, and the prediction intervals are pretty

Table 10.3: Our screening in comparison to other screenings using the proposed metrics defined by Marx and Lenhard (2011). With a Ratz index of 0.29, our approach to screening ranks in the lower midfield.

Screening		**Year**	**Test type**	**Duration**	**RATZ**	**PPV**	**Sensitivity**	**Specificity**	**Data**	**Dyslexic**
DP	Differenzierungs-probe	1975	Individual test	7 min.	25%	20%	33%	—	N=664	—
PB-LRS	Gruppentest zur Früherkennung von LRS	2004	Group test	60 min.	55%	36%	63%	87%	N=450	15%
RdH	Rundgang durch Hörhausen	2002	Individual test	45 min.	25% reading; 77% spelling	63%	38-48%	80%	N=375	20%
BISC	Bielefelder Screening	1999	Individual test	25 min.	—	50%	50%	—	—	15%
HASE	Heidelberger Auditive Screening in der Einschuluntersuchung	2007	Individual test	10 min.	49-59% depending on criterion	4-28%	69-76%	60-66	—	40%
WÜSC*	Würzburger Screening	2020	Individual test	25 min.	73%	54%	80%	83%	N=192	—
BS	Beuttler et al. Screening	2024	Group test	45 min.	46%	59%	59%	86%	N=272	29.10%/ 10.76%[a]

[a]In total 29.10% of the children were either weak in reading or writing, whereas 10.76% were weak in both skills.

Table 10.4: Predicted values and the according prediction intervals of selected children for both of their literacy skills.

Reading	predicted	lower	upper
$Child_a$	23.12	22.17	30.74
$Child_b$	32.49	26.14	44.62
$Child_c$	31.64	25.29	42.58

Spelling	predicted	lower	upper
$Child_a$	11.51	10.95	17.41
$Child_b$	11.44	10.84	16.23
$Child_c$	15.49	12.38	22.15

large. Nevertheless, the results are easy to read, provide a good basis for assessing the child's reading and spelling performance and include as much information about the prediction as possible, especially the amount of uncertainty.

10.4 Summary

In this chapter, we addressed several of our key research questions.

For *RQ2.1*, which focused on handling imbalances in the dataset, we employed the oversampling technique to generate synthetic test data for low-performing children. Additionally, we found that the final Random Forest model performed relatively well on imbalanced datasets, thanks to the aggregation of ensemble learning and its ability to separate outliers.

In response to *RQ2.2*, where we aimed to determine the most appropriate model for predicting literacy skills, we concluded that the Random Forest model outperformed other models, making it the most suitable choice for our predictive task. Its ability to integrate oversampling without compromising predictive power played a significant role in making this choice.

RQ2.3 focused on making our screening results comparable to other screenings. To achieve this, we had to classify the regression results, allowing us to calculate corresponding classification metrics. This led us to realize that the RATZ index for our screening approach was approximately 46%, placing it in the lower-middle range compared to other screenings. However, this classification step resulted in a loss of information, which might have negatively impacted the predictive performance of the screening process. To counteract this and answer *RQ2.4*, we proposed using estimation intervals, specifically predictive intervals. With this approach, we could predict specific reading or

spelling scores and intervals, which contain the target value with a certain probability (in our example 80%). This added level of uncertainty estimation made our regression results more informative and valuable for practical applications.

Overall, our research addressed these key questions, leading us to develop an effective Random Forest-based model for literacy screening while also exploring ways to improve interpretability through interval estimation techniques.

Part D

Overall Conclusion

Chapter 11

Summary

11.1 Evaluating feasibility and group testability

As we have argued, screenings in group settings are more cost and time-efficient than individually administered literacy screenings. Still, they are not always feasible, e.g., when they require the children to speak out loud or include complex tasks. We detailed the opportunities of digital technology to circumvent such problems: highly interactive tutorials with pedagogical agents, a balloon-popping mini-game at the end of each task, a symbol keypad system to start tasks together, a fantasy-themed environment, small rewards after each task, and simple navigation. The lessons we learned along the way and derived from our approach of designing such a group-based pre-reader screening are listed in Table 5.3. In general, we believe that the approach to playful assessment is particularly helpful for children with test anxiety, as the children perceived the screening and the tasks as games rather than tests.

The results further show that the screening was easy to use and that children perceived high usability, which indicates that the screening can be used independently by the children with minimal adult support (*RQ1.1*). We assume that the integration of highly interactive tutorials and straightforward navigation of the screening played crucial roles in achieving the feasibility of the screening for primary school children. The screening tasks were perceived very positively and considered games, and they did not invoke negative feelings, answering the second part of *RQ1.1*. Importantly, children with poor reading and spelling skills perceived the screening tasks similarly or even more favorable to typically developing children.

Finally, the screening proved to be feasible for group settings with up to 10 children for a single test administrator (*RQ1.2*), which we achieved with optimized seat arrangements, the use of headsets, and the implementation of (game) elements to engage children so they do not distract others while they perform the tasks.

Taken together, digital, game-based literacy screenings are feasible for group sessions. They leverage the advantages of digitization, such as automatic screening procedures, while being perceived positively, which might be particularly helpful for children with test anxiety or low literacy skills. We have introduced the term playful testing in the context of literacy screening, which consists of an embedded assessment, such as a stealth assessment, and includes a game-based approach. We showed that this approach motivates children to carry out the screening.

11.2 Findings of the screening study

To ensure consistency, we developed the tasks by incorporating the previously discovered insights from the user and game experience study, specifically emphasizing the underlying theoretical principles. The tasks were developed iteratively and frequently revised graphically and in terms of content. With this summary of the answer to *RQ1.3*, we would like to conclude the findings from the task evaluation and possible task improvements in the following chapters.

11.2.1 Task evaluation

In this thesis, we investigated several tasks to better understand their relationship to literacy skills in children. One of the tasks we explored was the Incidental Holistic Perception Task (IPET). In our investigation of the task, children's overall accuracy did not help to differentiate between children with weak and non-weak literacy skills, as their performance was near random level. Further, we couldn't confirm previous research findings on our pilot data, in which we observed significant recognition performance differences for weak literacy children between target and distractor items. However, we noticed interesting differences in item categories between real and fictitious items for the same children, especially when we looked at their response bias and accuracy.

In the Syllable Stress Task (SST), we found that both person parameters and sum scores were predictive of both literacy skills and the covariates Continuous Series Test (CST)

- our measure to indicate intelligence - and its interaction with age. Interestingly, there were no major differences between models using person parameters or sum scores to explain the literacy data, but the former tended to contribute slightly more to that. However, we expect this task to perform even better once the technical challenges are addressed to exclude children who act as if they are not interested in solving the task. The same analysis was done for the Reaction Time Discrimination Task (RTDT) task. We did not identify any predictive features in this task, neither through person parameters nor sum scores. The RTDT, like the SST task, showed relatively little explained variance for every model constructed to predict literacy skills. This is especially true for predicting spelling skills.

In the case of the Serial Reaction Time Task (SRTT), we discovered that accuracy was a good feature to distinguish performances between groups. At the same time, reaction times did not provide as much differentiation. Nonetheless, we found predictive value in modeling the learning curve using reaction times and calculating the individual learning curve, response increase and intercept coefficients, as it allowed us to conduct a more detailed analysis. This allows us to incorporate the effect of fatigue to some extent. However, we found that the task in general can be reduced in scope and complexity to battle task fatigue without losing predictive power.

Lastly, the Rapid Automatized Naming (RAN) task showed promising results in terms of prediction. As expected, the time needed per item proved to be a good predictor of reading and writing skills, and the number of wrong-naming errors was an additional predictive measure of spelling skills. For predicting the reading skills, the CST was a relevant covariate. Interestingly, compared to other tasks like the SRTT, there was no visible fatigue effect over the two pages of this task despite it being the last one to be completed. However, the manual postprocessing work for this task is extremely high compared to the other tasks.

In summary, tasks like SRTT and RAN, which have a well-established and proven concept behind them, and the SST, which we created from the ground up, show predictive characteristics. IPET and RTDT, on the other hand, produce less compelling results and require significant progress to be considered predictive of literacy skills. Because of the novel and exploratory implementation of the tasks, there is always a risk that the theoretically known features cannot be utilized efficiently. The results of the Principle Component Analysis (PCA) further suggest (1) that the predictors, and respectively our

tasks, cover a broad spectrum of proficiencies and thus help the screening to generalize well, and (2) that in principle, our eleven predictors can be reduced to a maximum of five components without losing too much information.

With this, we answered *RQ1.4* and refer to Table 7.18 for a complete overview of all predictors used for literacy skills prediction and further information about the PCA.

11.2.2 Recommendations for improving the tasks

IPET In contrast to the made-up objects from the original study by Hedenius et al. (2013), the fictitious animals in the IPET follow principles known from body composition. In the case of the animals, these are the head, body and legs. Accordingly, it is possible that these fictitious animals we create could not be perceived as fictional and therefore the task goal was altered. Likewise, the children seemed to guess a lot in the task, which indicates that the display duration was too short. The difficult task, however, is to find a short enough display time for the image grid to allow incidental recognition of the elements. We recommend that the assumptions made that children with weak literacy skills have better incidental and type-based perception be reexamined separately.

RTDT The task in its current form does not appear to contribute significantly to explaining the reading or spelling data. We have identified two possible causes for this:

1. The discriminatory power of the items should be increased. More significant group effects are expected due to greater differences in rise times. Modifying the task may also be advantageous so that stimuli from previous studies (e.g., Huss et al. (2011); Rauschenberger et al. (2018b)) can reduce bias or uncertainty from item generation.

2. Increased guessing parameters, especially with longer rise times, indicate that the task seems to be either too difficult for children of pre-reader age or that the auditory perception of the children with weak literacy skills is not as limited as assumed (Sauter et al., 2012).

SRTT The SRTT shows a fatigue effect in reaction times. The longer the exercise lasts and sequences are repeated, the longer the reaction times are, which is an opposite trend

to what is expected. As shown in Figure 7.5, the reaction time progression in blocks 2 to 4 is similar to the progression in the sequences within the blocks. We therefore suggest that the number of blocks and sequences inside the blocks can be reduced to counteract the fatigue effect.

11.3 Machine learning and its application in the literacy screening

In our study, we focused on predicting reading and spelling scores in children using learning algorithm models based on their task performances. Our goal was to predict the performance as accurately as possible with the information we had. For this, we modeled the task as a regression problem.

To facilitate our investigation, we designed a novel model training algorithm. To address imbalances in the dataset, we oversampled the areas of low-performing children, which is defined as one standard deviation below the mean. A custom oversampling implementation was necessary to prevent overfitting and data bleeding during the 5-fold cross-validation. The new training algorithm involved testing multiple models, multiple oversampling rates and assessing resulting model performances. We repeated this algorithm with different random seeds five times to ensure reliable and representative outcomes.

Our findings highlighted the Random Forest model's superiority in predicting reading and spelling skills. Notably, we achieved the best results when using an oversampling rate of 200% for children with lower reading and spelling performance levels. The Random Forest models incorporating oversampling exhibited notably enhanced predictive capabilities for children with weaker literacy skills, outperforming other models.

To make our screening process comprehensive and comparable, we classified the outcomes of our regression models for both literacy skills. This step demanded the implementation of custom methodologies to stratify and subsample based on two target variables. By classifying and evaluating the results from the regression data, we established a holistic and insightful perspective on the predictive power of our models, which proved to be medium to below average. By integrating prediction intervals, we have proposed a method to incorporate the uncertainty of the model in its prediction, making the results more precise and concrete to interpret.

Chapter 12

Outlook

The importance of language literacy early recognition systems is in full swing and is experiencing a new flourishing period, especially with the new technological developments in recent years (Endlich et al., 2024; Mancuso et al., 2024; Velmurugan, 2023). Looking ahead, the focus for a successful digital and game-based literacy screening tool for children of pre-reader age is on informative screening tasks, reliable data analysis and prediction using machine learning and integrating the screening within the educational systems.

On the one hand, it makes sense to revise the existing tasks and, if necessary, to evaluate them individually in further studies. Task duration should be significantly reduced to minimize cognitive load and fatigue effects, which, in retrospect, will most likely reduce unwanted bias in general. On the other hand, it is also worthwhile to discuss new tasks since phonological awareness tasks do not cover all phonological processes relevant to reading and spelling acquisition (Dodd, 2011; Stackhouse and Wells, 1997). The Speech Processing Model by Stackhouse and Wells (1997) provides a diagnostic framework to gain insight into relevant phonological competencies. The model describes different processes that are modularly involved in speech processing, like input, storage and output processes. The presented screening tool could be extended to incorporate the psycholinguistic model of speech processing to systematically measure the phonological competencies that seem fundamentally connected with the acquisition of written language. This would also allow for assessing the children's phonological competencies and enable customized training.

In this context and depending on the circumstances, it may also be worthwhile to collect

additional data and incorporate it into the screening, such as data from neuroimaging techniques. A recent review discusses how machine learning, specific models like artificial neural networks, support vector machines, and decision trees, has shown promise in identifying dyslexia through functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) data (Velmurugan, 2023). According to the authors, the literature they analyzed offers potential for early detection and personalized treatment, though further research is needed for validation and optimization. With the recent rise of artificial intelligence and large machine learning models, expanding the horizon for different and new technology seems very promising. However, gathering the additional data from fMRI or EEG has to be ethically discussed and is most likely very resource-intensive.

When calculating different literacy thresholds, this study and previous publications used the sample of native German speakers as a basis. In order to increase the generalizability of the screening, it may be advisable to assess the screenings and tasks' performance across diverse populations and cultural contexts. The goal is to ensure that the tool remains effective and unbiased for children from various backgrounds and language environments. In this context, it is also worth considering how such a screening system can be usefully implemented in existing educational institutions. A logical next step could be to intensify collaborations with educators and schools to understand how the tool can complement existing screening methods and provide valuable insights into early literacy development. A system that measures students' skills over their time in primary school and optimally in a stealth-assessment manner allows for more accurate analysis and prediction of skills. Therefore, we advise teachers and schools to conduct literacy screenings more often and use the game-based approach to integrate stealth assessment throughout primary school.

References

Abt, C. C. (1970). Serious Games. *American Behavioral Scientist*, 14(1):129–129.

Ameisen, E. (2018). Always start with a stupid model, no exceptions. https://blog.insightdatascience.com/always-start-with-a-stupid-model-no-exceptions-3a22314b9aaa.

Araújo, S. and Faísca, L. (2019). A Meta-Analytic Review of Naming-Speed Deficits in Developmental Dyslexia. *Scientific Studies of Reading*, 00(5):1–20.

Araújo, S., Reis, A., Petersson, K. M., and Faísca, L. (2015). Rapid automatized naming and reading performance: A meta-analysis. *Journal of Educational Psychology*, 107(3):868–883.

Barendregt, W. (2006). *Evaluating fun and usability in computer games with children*. PhD thesis, Department of Industrial Design.

Barry, W. J. (2003). Do Rhythm Measures Tell us Anything about Language Type? *15th ICPhS*, 1(November 2015):2693–2696.

Barth, K. and Gomm, B. (2004a). Gruppentest zur Früherkennung von Lese- und Rechtschreibeschwierigkeiten.

Barth, K. and Gomm, B. (2004b). *Gruppentest zur Früherkennung von Lese- und Rechtschreibschwierigkeiten: phonologische Bewusstheit bei Kindergartenkindern und Schulanfängern (PB-LRS)*. Number Bd. 2 in Reinhardt-Test. Reinhardt.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1).

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29.

Becker, N. (2021). The Right Way to Oversample in Predictive Modeling.

Beddington, J., Cooper, C. L., Field, J., Goswami, U., Huppert, F. A., Jenkins, R., Jones, H. S., Kirkwood, T. B., Sahakian, B. J., and Thomas, S. M. (2013). *From*

Stress to Wellbeing Volume 2, volume 2. Palgrave Macmillan UK, London.

Beddington, J., Cooper, C. L., Field, J., Goswami, U., Huppert, F. A., Jenkins, R., Jones, H. S., Kirkwood, T. B. L., Sahakian, B. J., and Thomas, S. M. (2008). The mental wealth of nations. *Nature*, 455(7216):1057–1060.

Bender, F., Brandelik, K., Jeske, K., Lipka, M., Löffler, C., Mannhaupt, G., Naumann, C. L., Nolte, M., Ricken, G., Rosin, H., Scheerer-neumann, G., and Aster, M. V. (2017). Die integrative Lerntherapie. 6(2):65–73.

Berkling, K. and Pflaumer, N. (2014). Phontasia - a Phonics Trainer for German Spelling in Primary Education. *Proceedings of the 4th Workshop on Child Computer Interaction (WOCCI 2014)*, (Wocci):33–38.

Bernstein, F. and Bernstein, F. H. (1998). *Ludi publici: Untersuchungen zur Entstehung und Entwicklung der öffentlichen Spiele im republikanischen Rom*. Number Band 119 in Historia Einzelschriften. Franz Steiner Verlag.

Berrar, D. (2018). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1-3(April):542–545.

Berrar, D., Thomas-Vaslin, V., Six, A., Bellier, B., and Klatzmann, D. (2013). Lymphocyte Labeling, Cell Division Investigation. In *Encyclopedia of Systems Biology*, chapter Overfittin, pages 1617–1619. Springer New York, New York, NY.

Bhattacharyya, S. (2018). Ridge and Lasso Regression: L1 and L2 Regularization. Towards Data Science.

Bocklet, T., Winterholler, C., Magnet, A., Maier, Schuster, M., and Noeth, E. (2009). An automatic screening test for preschool children: Theory and data collection.

Bowers, P. G. (1995). Tracing symbol naming speed's unique contributions to reading disabilities over time. *Reading and Writing*, 7(2):189–216.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., and Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education*, 94:178–192.

Bradley, L. and Bryant, P. E. (1983). Categorizing sounds and learning to read–a causal connection. *Nature*, 301(3):419–421.

Branco, P., Ribeiro, R. P., Torgo, L., Krawczyk, B., and Moniz, N. (2017). SMOGN: a Pre-processing Approach for Imbalanced Regression. *Proceedings of Machine Learning Research*, 74:36–50.

Brandelik, K. (2014). *Sprachrhythmische Fähigkeiten im Schriftspracherwerb*. Südwest-

deutscher Verlag für Hochschulschriften.

Breiman, L. (2001). *Random Forests*. PhD thesis, University of California, Berkeley.

Bressler, D. and Bodzin, A. (2013). A mixed methods assessment of students' flow experiences during a mobile augmented reality science game. *Journal of Computer Assisted Learning*, 29(6):505–517.

Brizzolara, D., Chilosi, A., Cipriani, P., Di Filippo, G., Gasperini, F., Mazzotti, S., Pecini, C., and Zoccolotti, P. (2006). Do Phonologic and Rapid Automatized Naming Deficits Differentially Affect Dyslexic Children With and Without a History of Language Delay? A Study of Italian Dyslexic Children. *Cognitive and Behavioral Neurology*, 19(3):141–149.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel Inference. *Sociological Methods & Research*, 33(2):261–304.

Catts, H. W., Gillispie, M., Leonard, L. B., Kail, R. V., and Miller, C. A. (2002). The Role of Speed of Processing, Rapid Naming, and Phonological Awareness in Reading Achievement. *Journal of Learning Disabilities*, 35(6):510–525.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6):1–29.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2016, pages 785–794, New York, NY, USA. ACM.

Chen, Y. C. (2017). Empirical study on the effect of digital game-based instruction on students' learning motivation and achievement. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(7):3177–3187.

Clark, D. B., Tanner-Smith, E. E., and Killingsworth, S. S. (2016). Digital Games, Design, and Learning. *Review of Educational Research*, 86(1):79–122.

Collection, L. C. (2018). German news corpus based on material crawled in 2018No Title.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., and Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers and Education*, 59(2):661–686.

Cordova, D. I. and Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88(4):715–730.

Corriveau, K., Pasquini, E., and Goswami, U. (2007). Basic Auditory Processing Skills and Specific Language Impairment: A New Look at an Old Hypothesis. *Journal of Speech, Language, and Hearing Research*, 50(3):647–666.

Corriveau, K. H. and Goswami, U. (2009). Rhythmic motor entrainment in children with speech and language impairments: Tapping to the beat. *Cortex*, 45(1):119–130.

Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., and Wood, F. B. (2006). Suicidality, School Dropout, and Reading Problems Among Adolescents. *Journal of Learning Disabilities*, 39(6):507–514.

Denckla, M. B. and Rudel, R. G. (1976). Rapid 'automatized' naming (R.A.N.): Dyslexia differentiated from other learning disabilities. *Neuropsychologia*, 14(4):471–479.

Deroost, N., Zeischka, P., Coomans, D., Bouazza, S., Depessemier, P., and Soetens, E. (2010). Intact first- and second-order implicit sequence learning in secondary-school-aged children with developmental dyslexia. *Journal of Clinical and Experimental Neuropsychology*, 32(6):561–572.

Derryberry, A. (2007). Serious games: online games for learning. Technical Report 9.

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). From game design elements to gamefulness. In *Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '11*, page 9, New York, USA. ACM Press.

Diehl, J. J., Frost, S. J., Sherman, G., Mencl, W. E., Kurian, A., Molfese, P., Landi, N., Preston, J., Soldan, A., Fulbright, R. K., Rueckl, J. G., Seidenberg, M. S., Hoeft, F., and Pugh, K. R. (2014). Neural correlates of language and non-language visuospatial processing in adolescents with reading disability. *NeuroImage*, 101:653–666.

Dilling, H., Mombour, W., Schmidt, M. H., and Organization, W. H. (2015). Internationale Klassifikation psychischer Störungen : ICD-10, Kapitel V; klinisch-diagnostische Leitlinien (10. Auflage).

Dirx, R. (1981). *Das Buch vom Spiel: das Spiel einst und jetzt*. Burckhardthaus-Verlag.

Djaouti, D., Alvarez, J., and Jessel, J.-P. (2011). Classifying Serious Games. pages 118–136.

Dodd, B. (2011). Differentiating Speech Delay From Disorder. *Topics in Language Disorders*, 31(2):96–111.

Dodd, B., Leahy, J., and Hambly, G. (1989). Phonological disorders in children: Underlying cognitive deficits. *British Journal of Developmental Psychology*, 7(1):55–71.

Dondlinger, M. (2007). Educational Video Game Design: A Review of the Literature. *Journal of applied educational technology*, 4(1):21–31.

Dooley, J. (2011). *Software Development and Professional Practice*. Apress, Berkeley, CA.

Dörner, D. and Bick, T. (1983). *Lohausen: vom Umgang mit Unbestimmtheit und Komplexität : [DFG-Projekt DO 200/4 "Systemdenken" Lehrstuhl Psychologie II der Universität Bamberg 1981]*. Huber.

Druin, A. (2002). The role of children in the design of new technology. *Behaviour & Information Technology*, 21(1):1–25.

Dummer-Smoch, L., Lehmann-Breuer, W., Ruoho, K., Hotulainen, R., Steinbrink, C., Klatte, M., and Lachmann, T. (2012). Kommentare zu Steinbrink, Schwanda, Klatte und Lachmann: Sagen Wahrnehmungsleistungen zu Beginn der Schulzeit den Lese-Rechtschreiberfolg in Klasse 1 und 2 voraus? (ZEPP, Heft 4/2010, S. 188–200). *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 44(2):92–105.

Dyslexia International (2014). Dyslexia International Report. Technical report, Brussels.

Einsiedler, W., Frank, A., Kirschhock, E.-M., Martschinke, S., and Treinies, G. (2002). Der Einfluss verschiedener Unterrichtsmethoden auf die phonologische Bewusstheit sowie auf Lese- und Rechtschreibleistungen im 1. Schuljahr. *Psychologie in Erziehung und Unterricht*, pages 194–209.

Endlich, D., Küspert, P., Lenhard, W., Marx, P., and Schneider, W. (2019). *LRS-Screening. Laute, Reime, Sprache - Würzburger Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten*. Hogrefe, Göttingen.

Endlich, D., Lenhard, W., Marx, P., and Richter, T. (2024). *Das Lese-Screening in LONDI -Konzeption, empirische Ergebnisse und praktischer Einsatz eines neuartigen Onlinescreenings für Leseschwierigkeiten*, pages 137–159.

Eseryel, D., Ge, X., Ifenthaler, D., and Law, V. (2011). Dynamic Modeling as a Cognitive Regulation Scaffold for Developing Complex Problem-Solving Skills in an Educational Massively Multiplayer Online Game Environment. *Journal of Educational Computing Research*, 45(3):265–286.

Esser, G., Wyschkon, A., and Schmidt, M. H. (2002). Was wird aus Achtjährigen mit einer Lese- und Rechtschreibstörung. *Zeitschrift für Klinische Psychologie und Psychotherapie*, 31(4):235–242.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018).

Learning from Imbalanced Data Sets. Springer International Publishing, Cham.

Forsa (2024). Die Schule aus Sicht der Schulleiterinnen und Schulleiter - Digitalisierung und digitale Ausstattung. Technical report, forsa Gesellschaft für Sozialforschung und statistische Analysen, Berlin.

Frank, A., Kirschhock, E.-M., and Martschinke, S. (2001). *Der Rundgang durch Hörhausen - Erhebungsverfahren zur phonologischen Bewusstheit Diagnose und Förderung im Schriftspracherwerb*. Auer, Donauwörth, 1 edition.

Frensch, P. A. and Rünger, D. (2003). Implicit Learning. *Current Directions in Psychological Science*, 12(1):13–18.

Freud, S. (1920). *Gesammelte Werke: chronologisch geordnet. Werke aus den Jahren 1917-1920. Zwölfter Band*. Imago Publishing.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gaggi, O., Galiazzo, G., Palazzi, C., Facoetti, A., and Franceschini, S. (2012). A Serious Game for Predicting the Risk of Developmental Dyslexia in Pre-Readers Children. In *2012 21st International Conference on Computer Communications and Networks (ICCCN)*, pages 1–5. IEEE.

Gaggi, O., Palazzi, C. E., Ciman, M., Galiazzo, G., Franceschini, S., Ruffino, M., Gori, S., and Facoetti, A. (2017). Serious games for early identification of developmental dyslexia. *Computers in Entertainment*, 15(2).

Galuschka, K., Ise, E., Krick, K., and Schulte-Körne, G. (2014). Effectiveness of treatment approaches for children and adolescents with reading disabilities: A meta-analysis of randomized controlled trials. *PLoS ONE*, 9(2).

Ganguin, S. (2010). *Computerspiele und lebenslanges Lernen: Eine Synthese von Gegensätzen*. Medienbildung und Gesellschaft. VS Verlag für Sozialwissenschaften.

Ge, X. and Ifenthaler, D. (2017). Designing Engaging Educational Games and Assessing Engagement in Game-Based Learning. pages 253–270. IGI Global, Hershey, PA.

Gebhard, C. (2012). *Sprechtempo im Sprachvergleich*. PhD thesis, Humboldt-Universität zu Berlin, Philosophische Fakultät II.

Gold, A. (2018). *Lernschwierigkeiten. Ursachen, Diagnostik, Intervention. 2., erweiterte und überarbeitete Auflage.* Kohlhammer Standards Psychologie. Verlag W. Kohlhammer, Stuttgart.

Good, P. I. and Hardin, J. W. (2008). *Common Errors in Statistics (and How to Avoid Them)*. Wiley-IEEE Press, third edition.

Goswami, U. (1999). The relationship between phonological awareness and orthographic representation in different orthographies. In *Learning to read and write: A cross-linguistic perspective.*, Cambridge studies in cognitive and perceptual development., pages 134–156. Cambridge University Press, New York, NY, US.

Goswami, U., Mead, N., Fosker, T., Huss, M., Barnes, L., and Leong, V. (2013). Impaired perception of syllable stress in children with dyslexia: A longitudinal study. *Journal of Memory and Language*, 69(1):1–17.

Hämäläinen, J. A., Salminen, H. K., and Leppänen, P. H. T. (2013). Basic Auditory Processing Deficits in Dyslexia. *Journal of Learning Disabilities*, 46(5):413–427.

Hanna, L., Neapolitan, D., and Risden, K. (2004). Evaluating computer game concepts with children. In *Proceeding of the 2004 conference on Interaction design and children building a community - IDC '04*, pages 49–56, New York, USA. ACM Press.

Hasselhorn, M., Schumann-Hengsteler, R., Gronauer, J., Grube, D., Mähler, C., Schmid, I., Seitz-Stein, K., and Zoelch, C. (2012). Arbeitsgedächtnistestbatterie für Kinder von fünf bis zwölf Jahren:(AGTB 5-12).

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d prime. *Behavior Research Methods, Instruments, & Computers*, 27(1):46–51.

He, H. and Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press.

Hedenius, M., Ullman, M. T., Alm, P., Jennische, M., and Persson, J. (2013). Enhanced recognition memory after incidental encoding in children with developmental dyslexia. *PloS one*, 8(5):e63998.

Hesketh, A., Dima, E., and Nelson, V. (2007). Teaching phoneme awareness to preliterate children with speech disorder: A randomized controlled trial. *International Journal of Language and Communication Disorders*, 42(3):251–271.

Hinderks, A., Schrepp, M., Rauschenberger, M., Olschner, S., and Thomaschewsik, J. (2012). Konstruktion eines Fragebogens für jugendliche Personen zur Messung der User Experience. *Usability Professionals Konferenz 2012*, pages 78–83.

Holz, H., Beuttler, B., Löfflad, D., and Ninaus, M. (2024). Developing a group-based literacy screening for german pre-readers: A digital, game-based approach. *Proc.*

ACM Hum.-Comput. Interact., 8(MHCI). In review.

Holz, H., Beuttler, B., and Ninaus, M. (2018). Design Rationales of a Mobile Game-Based Intervention for German Dyslexic Children. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '18 Extended Abstracts. ACM.

Holz, H., Ninaus, M., Schwerter, J., Parrisius, C., Beuttler, B., Brandelik, K., and Meurers, D. (2023). A digital game-based training improves spelling in German primary school children – A randomized controlled field trial. *Learning and Instruction*, 87:101771.

Howard, D. V. and Howard, J. H. (1989). Age differences in learning serial patterns: Direct versus indirect measures. *Psychology and Aging*, 4(3):357–364.

Hsu, H. J. and Bishop, D. V. (2014). Sequence-specific procedural learning deficits in children with specific language impairment. *Developmental Science*, 17(3):352–365.

Huizinga, J. (1955). *Homo ludens a study of the play-element in culture*. Beacon Press, Boston, Mass.

Huotari, K. and Hamari, J. (2012). Defining gamification. In *Proceeding of the 16th International Academic MindTrek Conference on - MindTrek '12*, page 17, New York, New York, USA. ACM Press.

Hura, S. and Echols, C. H. (1996). The role of stress and articulatory difficulty in children's early productions. *Developmental Psychology*, 32(1):165–176.

Huss, M., Verney, J. P., Fosker, T., Mead, N., and Goswami, U. (2011). Music, rhythm, rise time perception and developmental dyslexia: Perception of musical meter predicts reading and phonology. *Cortex*, 47(6):674–689.

Ifenthaler, D., Eseryel, D., and Ge, X. (2012). *Assessment in Game-Based Learning*. Springer New York, New York, NY.

Ifenthaler, D. and Kim, Y. J., editors (2019). *Game-Based Assessment Revisited*. Advances in Game-Based Learning. Springer International Publishing, Cham.

Ise, E., Engel, R. R., and Schulte-Körne, G. (2012). Was hilft bei der Lese-Rechtschreibstörung? *Kindheit und Entwicklung*, 21(2):122–136.

Ise, E. and Schulte-Körne, G. (2010). Spelling deficits in dyslexia: Evaluation of an orthographic spelling training. *Annals of Dyslexia*, 60(1):18–39.

Jansen, H., Mannhaupt, G., Marx, H., and Skowronek, H. (2002). *Bielefelder Screening zur Früherkennung von Lese-Rechtschreib- schwierigkeiten (BISC)*. Hogrefe Verlag, Göttingen.

Jiménez-Fernández, G., Gutiérrez-Palma, N., and Defior, S. (2015). Impaired stress awareness in Spanish children with developmental dyslexia. *Research in Developmental Disabilities*, 37:152–161.

Jung, S., Moeller, K., Klein, E., and Heller, J. (2021). Mode effect: An issue of perspective? Writing mode differences in a spelling assessment in German children with and without developmental dyslexia. *Dyslexia*, (December 2020):dys.1675.

Kant, I. and Rink, F. T. (1803). *Über Pädagogik*. Philosophische Bibliothek / Taschenausgaben. Nicolovius.

Kassambara, A. (2018). *Machine Learning Essentials: Practical Guide in R*. STHDA.

Kast, M., Baschera, G.-M., Gross, M., Jäncke, L., and Meyer, M. (2011). Computer-based learning of spelling skills in children with and without dyslexia. *Annals of Dyslexia*, 61(2):177–200.

Kiili, K., de Freitas, S., Arnab, S., and Lainema, T. (2012). The Design Principles for Flow Experience in Educational Games. *Procedia Computer Science*, 15:78–91.

Kiili, K. and Ketamo, H. (2018). Evaluating Cognitive and Affective Outcomes of a Digital Game-Based Math Test. *IEEE Transactions on Learning Technologies*, 11(2):255–263.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Koziol, L. F. and Budding, D. E. (2012). *Procedural Learning*, pages 2694–2696. Springer US, Boston, MA.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, Articles*, 28(5):1–26.

Kuhn, M. and Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman & Hall/CRC Data Science Series. CRC Press.

Landerl, K., Freudenthaler, H. H., Heene, M., De Jong, P. F., Desrochers, A., Manolitsis, G., Parrila, R., and Georgiou, G. K. (2019). Phonological Awareness and Rapid Automatized Naming as Longitudinal Predictors of Reading in Five Alphabetic Orthographies with Varying Degrees of Consistency. *Scientific Studies of Reading*, 23(3):220–234.

Laugwitz, B., Held, T., and Schrepp, M. (2008). Construction and evaluation of a

user experience questionnaire. *Proceedings of the 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society (USAB'08)*, 5298:63–76.

Learn, L. G. (2023). DORA - Dyslexia Screener for Schools. Website.

Leong, V., Hämäläinen, J., Soltész, F., and Goswami, U. (2011). Rise time perception and detection of syllable stress in adults with developmental dyslexia. *Journal of Memory and Language*, 64(1):59–73.

Lervåg, A. and Hulme, C. (2009). Rapid Automatized Naming (RAN) Taps a Mechanism That Places Constraints on the Development of Early Reading Fluency. *Psychological Science*, 20(8):1040–1048.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3):18–22.

Lim, S. and Reeves, B. (2010). Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human-Computer Studies*, 68(1-2):57–68.

Lindberg, S. (2016). Ursachen der Lese-Rechtschreibstörungen (LRS). In *Lese-Rechtschreibstörungen (LRS)*, Uni-Taschenbücher, chapter 4, pages 53–64. UTB GmbH.

Lovegrove, W., Martin, F., Bowling, A., Blackwood, M., Badcock, D., and Paxton, S. (1982). Contrast sensitivity functions and specific reading disability. *Neuropsychologia*, 20(3):309–315.

Lu, B. and Hardin, J. (2021). A Unified Framework for Random Forest Prediction Error Estimation. *Journal of Machine Learning Research*, 22(8):1–41.

Lum, J., Gelgic, C., and Conti-Ramsden, G. (2010). Procedural and declarative memory in children with and without specific language impairment. *International Journal of Language & Communication Disorders*, 45(1):96–107.

Lum, J. A., Ullman, M. T., and Conti-Ramsden, G. (2013). Procedural learning is impaired in dyslexia: Evidence from a meta-analysis of serial reaction time studies. *Research in Developmental Disabilities*, 34(10):3460–3476.

Macmillan, N. A. and Creelman, C. D. (1990). Response Bias: Characteristics of Detection Theory, Threshold Theory, and "Nonparametric" Indexes. *Psychological Bulletin*, 107(3):401–413.

Maimon, O. Z. and Rokach, L. (2014). *Data Mining With Decision Trees: Theory And Applications (2nd Edition)*. Series In Machine Perception And Artificial Intelli-

gence. World Scientific Publishing Company.

Mancuso, L., Tancredi, C., Provola, M., Marino, L., Coppola, E., and Presta, R. (2024). *Predicting Dyslexia in Children Through Game-Based Screening: Introducing Fluffy the Game*, pages 177–195.

Mannhaupt, G. (2006). *MÜSC Münsteraner Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten*. Cornelsen, Berlin.

Marx, H. (1992). Methodische und inhaltliche Argumente für und wider eine frühe Identifikation und Prädiktion von Lese-Rechtschreibschwierigkeiten. *Diagnostica*, 38(3):249–268.

Marx, P. and Lenhard, W. (2011). Diagnostische Merkmale von Screening-Verfahren zur Früherkennung möglicher Probleme beim Schriftspracherwerb. *Frühprognose schulischer Kompetenzen*, pages 68–84.

Marx, P. and Weber, J. (2006). Vorschulische Vorhersage von Lese- und Rechtschreibschwierigkeiten. *Zeitschrift für Pädagogische Psychologie*, 20(4):251–259.

Maurer, U., Bucher, K., Brem, S., and Brandeis, D. (2003). Altered responses to tone and phoneme mismatch in kindergartners at familial dyslexia risk. *NeuroReport*, 14(17):2245–2250.

Mayer, A. (2008). *Phonologische Bewusstheit, Benennungsgeschwindigkeit und automatisierte Leseprozesse. Aufarbeitung des Forschungsstandes und praktische Fördermöglichkeiten. 1. Auflage.* Berichte aus der Pädagogik. Shaker, Aachen.

Mayer, A. (2016). *Lese-Rechtschreibstörungen (LRS)*. Uni-Taschenbücher. UTB GmbH.

Mayringer, H. and Wimmer, H. (2014). *Salzburger Lese-Screening für die Schulstufen 2-9.* hogrefe.

Metelmann, L. (2020). *Die serial reaction time task (SRTT) als Prädiktor für Lese- und Rechtschreibschwäche bei Grundschüler*innen.* [Unpubulished bachelor thesis], Eberhard-Karls-Universität Tübingen.

Minuth, N. S. (2019). *Entwicklung und Evaluation einer Tablet-Applikation zur Messung des impliziten Lernens für die neuropsychologische Diagnostik.* Master's thesis, University of Tübingen.

Mohtaram, S., Che Pee, N., and Shibghatullah, A. (2017). DleksiaGame: A Mobile Dyslexia Screening Test Game to Screen Dyslexia Using Malay Language Instruction. *Asian Journal of Information Technology*, 16.

Moisello, C., Crupi, D., Tunik, E., Quartarone, A., Bove, M., Tononi, G., and Ghilardi, M. F. (2009). The serial reaction time task revisited: a study on motor sequence

learning with an arm-reaching task. *Experimental Brain Research*, 194(1):143–155.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307.

Moll, K., Fussenegger, B., Willburger, E., and Landerl, K. (2009). RAN is not a measure of orthographic processing. Evidence from the asymmetric German orthography. *Scientific Studies of Reading*, 13(1):1–25.

Moll, K., Kunze, S., Neuhoff, N., Bruder, J., and Schulte-Körne, G. (2014). Specific learning disorder: Prevalence and gender differences. *PLoS ONE*, 9(7).

Moll, K. and Landerl, K. (2009). Double Dissociation Between Reading and Spelling Deficits. *Scientific Studies of Reading*, 13(5):359–382.

Moll, K. and Landerl, K. (2010). *SLRT II – Lese- und Rechtschreibtest. Weiterentwicklung des Salzburger Lese- und Rechtschreibtests (SLRT)*. Verlag Hans Huber, Bern.

Moll, K., Wallner, R., and Landerl, K. (2012). Kognitive Korrelate der Lese-, Leserechtschreib- und der Rechtschreibstörung. *Lernen und Lernstörungen*, 1(1):7–19.

Moreno-Ger, P., Torrente, J., Hsieh, Y. G., and Lester, W. T. (2012). Usability Testing for Serious Games: Making Informed Design Decisions with User Data. *Advances in Human-Computer Interaction*, 2012:1–13.

Moschko, T. (2018). *Wiedererkennung nach beiläufigem Enkodieren – Erkundung kognitiver Vorteile lese-rechtschreibschwacher Grundschüler*. [Unpubulished bachelor thesis], Eberhard-Karls-Universität Tübingen.

Nau, R. (2021). How to compare models.

Neubrand, J. (2020). *Zum Einfluss der Wahrnehmung sprachlicher Betonungsmuster auf die Entwicklung der Leserechts chreibfertigkeiten bei Grundschüler*innen*. [Unpubulished bachelor thesis], Eberhard-Karls-Universität Tübingen.

Neugebauer, U. and Becker-Mrotzek, M. (2013). Die Qualität von Sprachstandsverfahren im Elementarbereich.

Nissen, M. J. and Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, 19(1):1–32.

Olive, D. J. (2007). Prediction intervals for regression models. *Computational Statistics & Data Analysis*, 51(6):3115–3122.

Ortiz, R., Estévez, A., Muñetón, M., and Domínguez, C. (2014). Visual and auditory perception in preschool children at risk for dyslexia. *Research in Developmental Disabilities*, 35(11):2673–2680.

Overy, K. (2000). Dyslexia, Temporal Processing and Music: The Potential of Music as an Early Learning Aid for Dyslexic Children. *Psychology of Music*, 28(2):218–229.

Parker, L. E. and Lepper, M. R. (1992). Effects of fantasy contexts on children's learning and motivation: Making learning more fun. *Journal of Personality and Social Psychology*, 62(4):625–633.

Parrila, R., Kirby, J. R., and McQuarrie, L. (2004). Articulation Rate, Naming Speed, Verbal Short-Term Memory, and Phonological Awareness: Longitudinal Predictors of Early Reading Development? *Scientific Studies of Reading*, 8(1):3–26.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12:2825–2830.

Permentier, M. (2004). *Historisches Wörterbuch der Pädagogik*. Beltz Handbuch. Beltz.

Piaget, J. (1975). *Nachahmung, Spiel und Traum: die Entwicklung der Symbolfunktion beim Kinde*. Gesammelte Werke (Studienausgabe). Klett-Cotta.

Poels, K., Ijsselsteijn, W., and Kort, Y. D. (2008). Development of the Kids Game Experience Questionnaire. In *Proceedings of Meaningful Play*.

Powell, D., Stainthorp, R., Stuart, M., Garwood, H., and Quinlan, P. (2007). An experimental comparison between rival theories of rapid automatized naming performance and its relationship to reading. *Journal of Experimental Child Psychology*, 98(1):46–68.

PwC (2020). Videospiele - Umsatzentwicklung des Videospielmarktes. https://www.pwc.de/de/technologie-medien-und-telekommunikation/german-entertainment-and-media-outlook-2019-2023.html.

Pykes, K. (2020). Oversampling and Undersampling.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radesky, J. S., Weeks, H. M., Ball, R., Schaller, A., Yeo, S., Durnez, J., Tamayo-Rios, M., Epstein, M., Kirkorian, H., Coyne, S., and Barr, R. (2020). Young Children's Use of Smartphones and Tablets. *Pediatrics*, 146(1).

Ramus, F. (2003). Developmental dyslexia: specific phonological deficit or general sensorimotor dysfunction? *Current Opinion in Neurobiology*, 13(2):212–218.

Rauschenberger, M., Baeza-Yates, R., and Rello, L. (2020a). Screening risk of dyslexia through a web-game using language-independent content and machine learning. pages 1–12.

Rauschenberger, M., Baeza-Yates, R., and Rello, L. (2020b). Screening risk of dyslexia through a web-game using language-independent content and machine learning. In *Proceedings of the 17th International Web for All Conference*, pages 1–12, New York, NY, USA. ACM.

Rauschenberger, M., Baeza-Yates, R., and Rello, L. (2022). A Universal Screening Tool for Dyslexia by a Web-Game and Machine Learning. *Frontiers in Computer Science*, 3(January):1–16.

Rauschenberger, M., Rello, L., and Baeza-Yates, R. (2018a). A tablet game to target dyslexia screening in pre-readers. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct - MobileHCI '18*, pages 306–312, New York, USA. ACM Press.

Rauschenberger, M., Rello, L., Baeza-Yates, R., and Bigham, J. P. (2018b). Towards language independent detection of dyslexia with a web-based game. *W4A '18: The Internet of Accessible Things*.

Rauschenberger, M., Rello, L., Baeza-yates, R., Gomez, E., and Bigham, J. P. (2017). Towards the Prediction of Dyslexia by a Web-based Game with Musical Elements. In *The Web for all conference Addressing information barriers - W4A*, pages 4–7.

Reents, M. (2020). *Überprüfung der Validität eines nonverbalen Intelligenzmaßes zur Diagnostik der Lese-rechtschreib-schwäche.* [Unpubulished bachelor thesis], Eberhard-Karls-Universität Tübingen.

Rello, L., Bayarri, C., Otal, Y., and Pielot, M. (2014). A computer-based method to improve the spelling of children with dyslexia. *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility - ASSETS '14*, pages 6–13.

Robertson, E. M. (2007). The Serial Reaction Time Task: Implicit Motor Skill Learning? *Journal of Neuroscience*, 27(38):10073–10075.

Rüsseler, J., Gerth, I., and Münte, T. F. (2006). Implicit Learning is Intact in Adult Developmental Dyslexic Readers: Evidence from the Serial Reaction Time Task and Artificial Grammar Learning. *Journal of Clinical and Experimental Neuropsy-*

chology, 28(5):808–827.

Sauter, K., Heller, J., and Landerl, K. (2012). Sprachrhythmus und Schriftspracherwerb. *Lernen und Lernstörungen*, 1(4):225–239.

Schmalz, X., Moll, K., Mulatti, C., and Schulte-Körne, G. (2019). Is Statistical Learning Ability Related to Reading Ability, and If So, Why? *Scientific Studies of Reading*, 23(1):64–76.

Schöler, H. and Brunner, M. (2008). *HASE Heidelberger Auditives Screening in der Einschulungsuntersuchung*. Hogrefe, Göttingen, 2 edition.

Schönweiss, F. (2007). *Münsteraner Rechtschreibanalyse (MRA)*. Lernserver : Interaktive Förderdiagnostik. Schönweiss.

Schrepp, M. (2018). *User Experience Questionnaire Handbook*.

Schrepp, M., Hinderks, A., and Thomaschewski, J. (2017a). Construction of a Benchmark for the User Experience Questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(4):40.

Schrepp, M., Hinderks, A., and Thomaschewski, J. (2017b). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(6):103.

Schulte-Körne, G. (2010). The Prevention, Diagnosis, and Treatment of Dyslexia. *Deutsches Ärzteblatt international*.

Schulte-Körne, G. (2015). Diagnostik und Behandlung von Kindern und Jugend- lichen mit Lese- und / oder Rechtschreibstörung. *Deutsche Gesellschaft für Kinder und Jugendpsychiatrie, Psychosomatik und Psychotherapie e.V.*

Schulte-Körne, G. and Remschmidt, H. (2003a). Legasthenie – Symptomatik, Diagnostik, Ursachen, Verlauf und Behandlung. *Deutsches Ärzteblatt*, 100(7):396–406.

Schulte-Körne, G. and Remschmidt, H. (2003b). Legasthenie – Symptomatik, Diagnostik, Ursachen, Verlauf und Behandlung. *Deutsches Ärzteblatt*, 100(7):396–406.

Shute, V. and Levy, R. (2009). Assessment and learning in intelligent educational systems: A peek into the future. *... in Education ...*, (January):1–10.

Shute, V. and Ventura, M. (2013a). Stealth Assessment. In *Stealth Assessment: Measuring and Supporting Learning in Video Games*. The MIT Press.

Shute, V. and Ventura, M. (2013b). Stealth Assessment.

Shute, V. J. (2014). Stealth Assessment in computer-based games to support learning. Computer G(January 2011):503–523.

Snow, C. and Jones, J. (2001). Making a Silk Purse. *Education Week*, 20(32):60.

Snowling, M. J. (1995). Phonological processing and developmental dyslexia. *Journal of Research in Reading*, 18(2):141–423.

Sprenger-Charolles, L., Siegel, L. S., Jiménez, J. E., and Ziegler, J. C. (2011). Prevalence and reliability of phonological, surface, and mixed profiles in dyslexia: A review of studies conducted in languages varying in orthographic depth. *Scientific Studies of Reading*, 15(6):498–521.

Stackhouse, J. and Wells, B. (1997). *Children's speech and literacy difficulties: a psycholinguistic framework*. Singular Pub. Group, San Diego, California.

Stanat, P., Schipolowski, S., Schneider, R., Weirich, S., Henschel, S., and Sachse, K. (2023). *IQB-Bildungstrend 2022. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich*.

Stein, J. (2001). The magnocellular theory of developmental dyslexia. *Dyslexia*, 7(1):12–36.

Steinbrink, C., Schwanda, S., Klatte, M., and Lachmann, T. (2010). Sagen Wahrnehmungsleistungen zu Beginn der Schulzeit den Lese-Rechtschreiberfolg in Klasse 1 und 2 voraus? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 42(4):188–200.

Stoodley, C. J., Ray, N. J., Jack, A., and Stein, J. F. (2008). Implicit Learning in Control, Dyslexic, and Garden-Variety Poor Readers. *Annals of the New York Academy of Sciences*, 1145(1):173–183.

Tafti, M. A., Hameedy, M. A., and Baghal, N. M. (2009). Dyslexia, a deficit or a difference: Comparing the creativity and memory skills of dyslexic and nondyslexic students in Iran. *Social Behavior and Personality: an international journal*, 37(8):1009–1016.

Thomson, J. M. and Goswami, U. (2008). Rhythmic processing in children with developmental dyslexia: Auditory and motor rhythms link to reading and spelling. *Journal of Physiology-Paris*, 102(1-3):120–129.

Thornton, G. C. and Cleveland, J. N. (1990). Developing managerial talent through simulation. *American Psychologist*, 45(2):190–199.

Torgo, L. (2017). Resampling Strategies for Regression.

Torgo, L. and Ribeiro, R. (2007). Utility-Based Regression. In *Knowledge Discovery in Databases: PKDD 2007*, pages 597–604. Springer Berlin Heidelberg, Berlin, Heidelberg.

Treutlein, A., Roos, J., and Schöler, H. (2011). Kreuzvalidierung des Heidelberger Au-

ditiven Screenings in der Einschulungsuntersuchung (HASE). page 15.

Tröster, H., Flender, J., Reineke, D., and Wolf, S. M. (2004). *Dortmunder Entwicklungsscreening für den Kindergarten (DESK 3-6)*. Hogrefe, Göttingen.

Tröster, H., Flender, J., Reineke, D., and Wolf, S. M. (2016). *Dortmunder Entwicklungsscreening für den Kindergarten (DESK 3-6) – Revision*. Hogrefe, Göttingen.

Ulrich, R. and Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, 123(1):34–80.

Van Belle, G. (2008). *Statistical Rules of the thumb*. Wiley-IEEE Press, second edition.

van der Kleij, S. W., Groen, M. A., Segers, E., and Verhoeven, L. (2019). Sequential Implicit Learning Ability Predicts Growth in Reading Skills in Typical Readers and Children with Dyslexia. *Scientific Studies of Reading*, 23(1):77–88.

Velmurugan, S. (2023). Predicting Dyslexia with Machine Learning: A Comprehensive Review of Feature Selection, Algorithms, and Evaluation Metrics. *Journal of Behavioral Data Science*, 3(1):1–14.

Vicari, S., Finzi, A., Menghini, D., Marotta, L., Baldi, S., and Petrosini, L. (2005). Do children with developmental dyslexia have an implicit learning deficit? *Journal of Neurology, Neurosurgery & Psychiatry*, 76(10):1392–1397.

Vicari, S., Marotta, L., Menghini, D., Molinari, M., and Petrosini, L. (2003). Implicit learning deficit in children with developmental dyslexia. *Neuropsychologia*, 41(1):108–114.

Vidyasagar, T. R. and Pammer, K. (2010). Dyslexia: a deficit in visuo-spatial attention, not in phonological processing. *Trends in Cognitive Sciences*, 14(2):57–63.

von Károlyi, C. (2001). Visual-Spatial Strength in Dyslexia. *Journal of Learning Disabilities*, 34(4):380–391.

Von Károlyi, C., Winner, E., Gray, W., and Sherman, G. F. (2003). Dyslexia linked to talent: Global visual-spatial ability. *Brain and Language*, 85(3):427–431.

Wagner, R. K. and Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, 101(2):192–212.

Wainer, H., Gessaroli, M., and Verdi, M. (2006). Visual Revelations. *CHANCE*, 19(1):49–52.

Wimmer, H. (1993). Characteristics of developmental dyslexia in a regular writing system. *Applied Psycholinguistics*, 14(1):1–33.

Wolf, K. M., Schroeders, U., and Kriegbaum, K. (2016). Metaanalyse zur Wirk-

samkeit einer Förderung der phonologischen Bewusstheit in der deutschen Sprache. *Zeitschrift für Pädagogische Psychologie*, 30(1):9–33.

Wolf, M., Bowers, P. G., and Biddle, K. (2000). Naming-Speed Processes, Timing, and Reading. *Journal of Learning Disabilities*, 33(4):387–407.

Wolf, M., O'Rourke, A. G., Gidney, C., Lovett, M., Cirino, P., and Morris, R. (2002). The second deficit: An investigation of the independence of phonological and naming-speed deficits in developmental dyslexia. *Reading and Writing*, 15:43–72.

Wölfl, V. (2020). *Wiedererkennung nach beiläufigem Enkodieren – Erkundung kognitiver Vorteile lese-rechtsasdasdchreibschwacher Grundschüler.* [Unpubulished bachelor thesis], Eberhard-Karls-Universität Tübingen.

Zaman, B. (2008). Introducing contextual laddering to evaluate the likeability of games with children. *Cognition, Technology & Work*, 10(2):107–117.

Zhang, Z. (2016). Decision tree modeling using R. *Annals of Translational Medicine*, 4(15):275–275.

Zichermann, G. and Cunningham, C. (2011). *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps.* O'Reilly Media, Inc., 1st edition.

Appendix A

Screening Tasks

A.1 Incidental Holistic Perception Task (IPET)

Tutorial

In the following, translated instructions from german to english are given in quotation marks and the procedure is displayed in brackets:
(First blue agents (B) jumps into spotlight, then the red agent (R)) (B) "Hey!" (R) "Hey! Look!" (As in a game trial, the presentation of the grid image follows. First the fixation cross, then the grid image and finally the masking. Timings are the same as in a game trial) (B) "Wow, that was fast. Could you see what there was to see?" (A figure for recognition appears between the agents) (R) "Do you think you just saw this figure? If so, tap the green button. If you did not see it, tap the red button." (Waiting for input. The tutorial continues the same regardless of which button was pressed) (R) "This is fun. Do it again." (Presentation of the grid image with fixation cross and masking. After that a figure appears in the middle of the screen) (R) "Do you think you have just seen this figure?". (Waiting for input) (R) "Cool, one more!" (A second figure for recognition appears. Waiting for input) (R) "Yes, just like that!" (B) "I think now you know how to do it." (R) "Let's get really started." (End of tutorial)

Table A.1: All 12 items used in the Syllable Stress Task. The degree of difficulty of the distraction patterns (category 1-3) originally created by a language expert can be put into perspective using the Hamming and Levenshtein distance values.

Item	Sentence	Target pattern	Distractor pattern	Category	Hamming dist.	Levenshtein dist.
1	Ich renne und hüpfe	dDddDd	DdddD	1	-	3
2	Mäuse wühlen	DdDd	DddDd	1	-	3
3	Ole kann gut tauchen	DdddDd	dDddD	1	-	3
4	Die Kinder spielen	dDdDd	ddDdDd	1	-	5
5	Clara hat Fieber	DddDd	dDddD	2	4	4
6	Opa hat Geburtstag	DdddDd	dDdddD	2	4	4
7	In der Wüste sind Schlangen	ddDdddDd	ddDdDDD	3	4	2
8	In den Bergen gibt es Schnee	ddDdddD	ddDDdDd	2	2	3
9	Ich backe Kuchen	dDdDd	dDddD	3	3	2
10	Krokodile sind gefährlich	ddDdddDd	dDddddDd	3	2	2
11	Wir fahren in den Urlaub	dDdddDd	dDddDdd	3	2	2
12	Tamara ist Schlittenfahren	dDddDddd	dDddDddD	3	1	1

A.2 Syllable Stress Task (SST)

A.2.1 Tutorial

In the following, translated instructions from german to english are given in quotation marks and the procedure is displayed in brackets:
(Yellow agent jumps into frame) "Hey, look!" (The yellow agent jumps aside and a picture with a speaker icon next to it appears in the middle of the screen. A female speaker says: The fir tree is green) "Here comes a sound pattern, listen! (The stress pattern to the said sentence sounds) "Now comes a dot pattern, look!". (The dot pattern appears) "Now watch this. The sound pattern matches the dot pattern." (The stress pattern sounds again and parallel to this, the respective syllables in the dot pattern are marked) "And the tone pattern also fits the sentence well" (The sentence is repeated again, at the same time the stress pattern is played and the dots in the dot pattern are highlighted) "That sounds like I'm singing. When you sing, sentences also match tones. Whenever a dot pattern matches the sentence, click on the dot pattern. Go ahead!" (A hand symbol appears on the dot pattern. Waiting for input) "Now click on the arrow key. Then continue." (The scene is cleaned up and the next image with a speaker icon next to it appears in the middle of the screen. A speaker says: The children are in the swimming pool) "Attention, now two sound patterns are coming. (The picture is shrinking and two dot patterns appear one after the other below the picture. When they appear, the corresponding stress patterns are played) "And here comes your task: Only one tone pattern fits the sentence well. The other does not fit so well. Your task is to find out which tone pattern fits the sentence best. You can listen to everything again by pressing these buttons. Try it out!" (The speaker symbols are overlaid with a hand symbol and highlighted. Waiting for input) "Well, do you already know the solution?" (A speaker says: "The children are in the swimming pool. At the same time the first dot pattern is highlighted in red. Waiting for input) "Exactly, tone pattern one fits. (Scene is cleaned up) "Now you can try the next two sentences all by yourself." (Next trial starts without additional instructions with the sentence: Goats live in the mountains. Waiting for input) "Great, one last time, then it really starts." (Next trial starts without additional instructions with the sentence: "In summer the sun is shining. Waiting for input) "You are doing great! Let's go!" (End of the tutorial)

A.3 Rise Time Discrimination Task (RTDT)

A.3.1 Tutorial

In the following, translated instructions from german to english are given in quotation marks and the procedure is displayed in brackets:
(Yellow agent jumps into scene, to the middle of the screen) "Hey, listen closely!" (A sound is played and individual note symbols appear) "The sound is from a dragon calling out of the forest. Look, there he is." (A dragon flies in on the left side and makes the same sound) "Click on the dragon and then on the arrow key. Then go on" (A symbol of a hand appears on the dragon. When the dragon is clicked, the button with the arrow symbol becomes active. Wait for input from the child) "Excellent, go on. Listen again." (A sound is played and individual note symbols appear) "There was a dragon calling again. But look, now two dragons come out of the forest" (Two dragons fly in, one on each side. Each dragon makes a sound after landing) "Which one has just called out of the forest? You can listen to it again by pressing these buttons. Try it!" (The speaker symbols under the dragons and in the forest are activated and overlaid with a hand symbol) "Well, do you know the solution yet?" (The dragon on the right side is activated) "This one called out of the forest. Tap him and then tap next" (Waiting for the child to enter) "Great, now you can start! (End of the tutorial)

A.4 Serial Reaction Time Task (SRTT)

A.4.1 Pseudorandomized sequence

NORTH, SOUTH, NORTH, WEST, SOUTH, EAST, WEST, EAST, NORTH, EAST, NORTH, EAST, NORTH, WEST, EAST, WEST, SOUTH, NORTH, SOUTH, EAST, NORTH, SOUTH, EAST, NORTH, EAST, WEST, SOUTH, NORTH, WEST, EAST, NORTH, WEST, EAST, NORTH, EAST, WEST, SOUTH, NORTH, SOUTH, EAST, NORTH, SOUTH, EAST, WEST, OUTH, NORTH, EAST, NORTH, WEST, EAST

A.4.2 Tutorial

In the following, translated instructions from German to English are given in quotation marks and the procedure is displayed in brackets:
(Stimulus is at position four) "Hello hello are you playing catch with me? I always jump

		Left Position				Correct		Right Position			
Item	Category	Rise Time	Steady State	Fall Time	Length			Rise Time	Steady State	Fall Time	Length
T1	1	15	735	50	800	x		—	—	—	—
T2	1	15	735	50	800		x	300	640	50	990
1	2	300	545	50	895		x	110	545	50	705
2	1	15	735	50	800		x	205	545	50	800
3	1	300	450	50	800	x		110	640	50	800
4	1	15	735	50	800		x	110	640	50	800
5	1	110	640	50	800		x	300	450	50	800
6	1	15	735	50	800	x		300	450	50	800
7	2	205	640	50	895	x		300	640	50	990
8	2	300	735	50	1085	x		15	735	50	800
9	1	110	640	50	800		x	205	545	50	800
10	1	205	545	50	800	x		300	450	50	800
11	2	15	735	50	800		x	300	735	50	1085
12	2	205	450	50	705	x		110	450	50	610
13	2	110	545	50	705	x		300	545	50	895
14	1	205	545	50	800	x		15	735	50	800
15	2	110	450	50	610		x	205	450	50	705
16	1	300	450	50	800	x		15	735	50	800
17	2	205	640	50	895		x	300	640	50	990
18	1	110	640	50	800	x		15	735	50	800
19	1	110	640	50	800		x	205	545	50	800
20	1	205	545	50	800		x	300	450	50	800

Table A.2: Item set of RTDT task with two training items

from one white square to another". (Jumps once into each field) "When I jump into a field, tap me as fast as you can. But use only one hand." (Jumps once into each field, but now always into the next one only after it has been tapped. The corresponding field flashes red) "Super! Try it again. Now the fields no longer blink." (Same procedure, only without flashing fields) "Yay! I think you have understood the task. Click on the arrow. Then we'll really get started." (Arrow appears. As already known from the previous tasks, the arrow changes color first to red, then to yellow, then to green after it is tapped. Then the task starts. Ende des Tutorials)

A.5 Rapid Automatized Naming Task (RAN)

A.5.1 Tutorial

In the following, translated instructions from german to english are given in quotation marks and the procedure is displayed in brackets:
(Blue agent jumps into frame) "Look. I brought us tickets." (All cards are uncovered and slowly read out by the blue agent) "Tree. Ball. Ice. Fish. Heart." (Cards are faced down and a shuffle sound appears) "I have now shuffled the cards and covered them. Pay attention, I'm going to read out all the cards as fast as I can. To get started you have to give the start signal. Press the start button." (Agent gets smaller and moves to the side. Tutorial continues after child pressed start) "Heart. Fish. Tree. Heart. Ice. ball. tree. Ice. fish. heart. ball. tree. Ice. Fish. tree. Heart. ball. Ice. tree. fish." (A hand symbol appears above the stop button, which, together with a sound effect, implies that the trial is finished) "Did you pay attention? After the last card I quickly pressed the stop button. Now it's your turn to read aloud. When you are ready, press the start button. Once the cards are revealed, read them to me as fast as you can." (A training trial follows) "Great, that's it. Now you can do two pages in a row." (End of the tutorial)

A.6 User Experience Questionnaire (UEQ)

Table A.3: All items from the UEQ questionnaire in German assigned to the categories.

User Experience Questionnaire(UEQ)	*Items*	
Attractiveness	12:	gut – schlecht
	16:	unangenehm – angenehm
	24:	schön – hässlich
Dependability / Ease of Use	11:	schwer zu bedienen – leicht zu bedienen
Perspicuity	2:	unverständlich – verständlich
	4:	leicht zu lernen – schwer zu lernen
	13:	kompliziert – einfach
	21:	übersichtlich – verwirrend
Simulation	5:	erfrischend – einschläfernd
	6:	langweilig – spannend
	7:	uninteressant – interessant
	18:	abwechslungsreich – eintönig

A.7 Measures and logging

Table A.4: Description of all variables logged from tablet data (1/2)

Task	Name of variable	Description
All	ProbandenID	Children-specific unique code of the child associated with the trial
	ReactionTime	Time in ms it took for the child to complete the trial
	Result	Validated input for each child. 1 = input was correct, 0 = input was wrong
	TaskNo, SubTask	TaskNo marks the numbering of the trial. If the task has a subtask, SubTask marks the numbering of the trial instead.
IPET	Task	Edited set
	Answer	Given answer by the child. True marks animal as target (green button), False marks it as a distractor (red button)
	CorrectAnswer	Is the displayed animal a target (True) or a distractor (False)
	PartialImg	The name of the animal image file
SST	FirstActionRT	Time in ms from the start of the task to the first registered input of the child
	IntroAnimationFinished	Time in ms from the start of the task to the end of all presentation animation. Input is being enabled at this point
	ListenCountCorrectSP	The number of clicks that were made on the button of the target stimulus to repeat the audio file
	ListenCountDistractorSP	The number of clicks that were made on the button of the distractor stimulus to repeat the audio file
	ListenCountMainImg	The number of clicks that were made on the button of the image representing the sentence to repeat the audio file
	Sentence	The current active sentence as string
	CorrectSP	The correct stress pattern corresponding to the displayed sentence (d = unstressed, D = stressed)
	DistractorSP	The stress pattern that functions as a distractor (d = unstressed, D = stressed)
	SelectedSP	The stress pattern selected by the child (d = unstressed, D = stressed)
RTDT	TaskNo	Trial number
	FirstActionRT	Time in ms from the start of the task to the first registered input of the child
	IntroAnimationRT	Time in ms from the start of the task to the end of all presentation animation. Input is being enabled at this point
	CorrectSoundFile	Name of the sound file of the target stimulus. The name includes information about rise time and steady state time of this simulus
	cPos	Position of the target stimulus. r = right side, l = left side
	DistractorSoundFile	Name of the sound file of the distractor stimulus. The name includes information about rise time and steady-state time of this simulus
	dPos	Position of the distractor stimulus. r = right side, l = left side
	SelectedSoundFile	Name of the selected stimulus sound file. The name contains rise time and steady-state duration
	Pressed	Position of the pressed stimulus, which is registered as input. r = right side, l = left side
	ListenCount	The number of clicks that were made on the button of the target stimulus to repeat the audio file
	ListenCountDistractorDragon	The number of clicks that were made on the button of the distractor stimulus to repeat the audio file
	ListenCountForestSound	The number of clicks that were made on the button of the repeat symbol to repeat the audio file

Table A.5: Description of all variables logged from tablet data (2/2)

Task	Name of variable	Description
SRTT	TaskNo	Block number
	SubtaskNo	Sequence number
	ReactionTime	Time it took for the child to press the next stimuli in ms
RAN	Task, TaskNo	Displays the page number
	T0ResetSW	Timestamp when the stopwatch's reset occurs (i.e., as soon as the RAN scene is loaded and the stopwatch is set to 0)
	T1StartClick	Time in ms that has passed since t0. Registered when the child clicks the start button
	T2RevealCard	Time in ms that has passed since t0. Registered when the symbols are revealed after the countdown
	T3StopClick	Time in ms that has passed since t0. Registered when the child clicks the stop/finish button
	T4EndRecording	Time in ms that has passed since t0. Registered when the system successfully saved the audio file
	FileLength.ms	Length of the recorded audio file in ms
	EditedAudioLength.s	Length of the manually edited audio file in seconds
	ErrorCount	Total count of all errors made in this trial, manually evaluated
	ErrorType1	Type of error made, manually evaluated
	ErrorType2	Type of error made, manually evaluated

Appendix B

Screening Methods and Results

B.1 Subject description

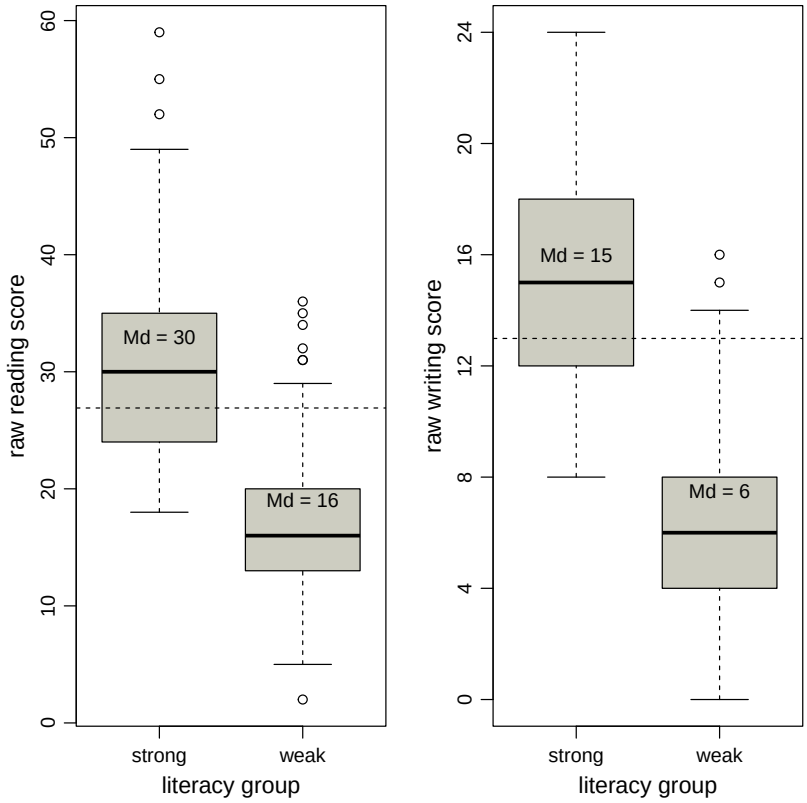


Figure B.1: Raw literacy skills between both literacy skill groups. The data is based on all children who have registered and completed at least the reading and writing tests ($n = 409$).

B.2 Task exclusions

Table B.1: IPET specific exclusions

participants after general exclusion	414
excludable behavior was observed	-9
missing literacy values	-5
participants left for analysis	400

Table B.2: SST specific exclusions

participants after general exclusion	414
excludable behavior was observed	-6
missing literacy values	-5
participants left for analysis	403

Table B.3: RTDT specific exclusions

participants after general exclusion	414
excludable behavior was observed	-10
systematic behavior[22]	-34
median reaction time $<$200 ms	-4
missing literacy values	-5
participants left for analysis	361

[22]Systematic behavior is the constant clicking only on one side or the constant alternation of the sides.

Table B.4: SRTT specific exclusions

participants after general exclusion	414
excludable behavior was observed	-26
>50% incorrect trials (scenario 1)[23]	-18
missing literacy values	-5
participants left for analysis	365

Table B.5: RAN specific exclusions

participants after general exclusion	414
faulty data[24]	-64
missing literacy values	-4
participants left for analysis	346

[23]Scenario 1 includes the exclusion of children that have answered 50% of all trials incorrectly.
[24]Faulty data refers to defective audio recordings and incomplete performance of the task.

B.3 Task methods and results

B.3.1 SST and RTDT IRT Models

Table B.6: IRT models considered and their associated item characteristic functions. The visualization is based on Neubrand (2020) and Wölfl (2020). $P(U_{ij})$ indicates the likelihood of person a_i's response in task x_j (with $U_{ij} = 1$ if the task is solved, $U_{ij} = 0$ if the response is incorrect).

Model	Item Characteristic Function
Rasch	$P(U_{ij} = u_{ij}\|\Theta_i, b_j) = \frac{exp[u_{ij}(\Theta_i - b_j)]}{1 + exp(\Theta_i - b_j)}$ with person parameter Θ_i and item difficulty b_j
Rasch with guessing	$P(U_{ij} = u_{ij}\|\Theta_i, a_j, b_j, g_j) = u_{ij}g_j + (1 - g_j)\frac{exp[u_{ij}a_j(\Theta_i - b_j)]}{1 + exp[a_j(\Theta_i - b_j)]}$ with person parameter Θ_i, item difficulty b_j, guessing parameter g_j and equalized discriminatory power a_j for all items
3PL	$P(U_{ij} = u_{ij}\|\Theta_i, a_j, b_j, g_j) = u_{ij}g_j + (1 - g_j)\frac{exp[u_{ij}a_j(\Theta_i - b_j)]}{1 + exp[a_j(\Theta_i - b_j)]}$ with person parameter Θ_i, item difficulty b_j, guessing parameter g_j and discriminatory power a_j
4PL	$P(U_{ij} = 1\|\Theta_i, a_j, b_j, g_j, u_j) = g_j + (u_j - g_j)\frac{exp[u_{ij}a_j(\Theta_i - b_j)]}{1 + exp[a_j(\Theta_i - b_j)]}$ with person parameter Θ_i, item difficulty b_j, guessing parameter g_j, discriminatory power a_j and maximum achievable probability of success u_j

B.3.2 SRTT results

Table B.7: Z-statistics of error rates between all blocks compared with bonferroni-corrected p-values of the SRTT.

contrast	estimate	SE	df	z.ratio	p.value
1 - 2	-0.28	0.13	Inf	-2.19	.287
1 - 3	-0.78	0.13	Inf	-6.00	**<.001**
1 - 4	-0.64	0.13	Inf	-4.83	**<.001**
1 - 5	-0.77	0.14	Inf	-5.53	**<.001**
2 - 3	-0.49	0.10	Inf	-5.03	**<.001**
2 - 4	-0.36	0.12	Inf	-2.98	**.029**
2 - 5	-0.49	0.12	Inf	-4.16	**<0.001**
3 - 4	0.13	0.10	Inf	1.40	1.000
3 - 5	0.01	0.10	Inf	0.06	1.000
4 - 5	-0.13	0.10	Inf	-1.35	1.000

Table B.8: Pairwise comparison of the blocks based using wilcoxon rank sum test for dependent samples with a bonferroni correction. All blocks differ significantly from block five in terms of mean reaction time. No other significant differences can be identified.

block	V	p
1-2	27219	.810
1-3	26784	.938
1-4	30305	.467
1-5	15864	**<.001**
2-3	27274	.788
2-4	27320	.768
2-5	14700	**<.001**
3-4	26685	.612
3-5	14115	**<.001**
4-5	11816	**<.001**

B.3.3 RAN Results

Table B.9: Stepwise reduction of reading prediction model with covariates age, gender and CST for the RAN task, starting with the most complex model. Each step, the predictor with least evidence (p) is dropped until only predictors, that contribute significantly in reading skill prediction, are left.

Predictors	1. Step			2. Step			3. Step		
	t value	p	drop	t value	p	drop	t value	p	drop
Age	-0.11	.912		-0.08	.939		-0.10	.922	
Gender	-0.33	.740		-0.32	.747		-0.33	.741	
CST	1.91	.057		1.95	.052		1.98	**<.048**	
Itemtime	-7.90	**<.001**		-8.10	**<.001**		-8.19	**<.001**	
Wrong-naming	-1.23	.218		-1.33	.184	x			
Omitted-word	-0.76	.449	x						
Age:gender	0.38	.707		0.37	.715		0.37	.713	
Age:CST	-1.59	.111		-1.63	.104		-1.67	.095	
Gender:CST	0.03	.978		0.02	.988		0.01	.990	
Age:gender:CST	-0.07	.943		-0.06	.953		-0.05	.963	

Appendix C

Predictive Data Analysis

C.1 Models considered for model training

Our model selection was primarily based on regression models that have been used in similar research settings, models that are widely used for regression problems and documented in the machine learning environment, and models that are optimally well supported in R. With Rauschenberger et al. (2020b)'s research, we found a project with similar goals that also employs machine learning technologies to predict children's future reading and writing development. Further models could be identified, among others, via the scikit-learn framework, which are specifically aligned for the programming language python, but contain good explanations of the common learning algorithm models. Table C.1 shows all models of learning algorithms considered for this thesis and what hyperparameter are used for tuning while training the models.

Table C.1: Listing of all models considered for training and their hyperparameters, if any. For each model with hyperparameters, a grid was created with all possible parameter combinations and the best combination was determined when training the model.

Models	Hyperparameter	Description
Dummy Regressor	-	-
Linear Regression (lm)	-	-
rPart	maxdepth (1:25; 1)	Maximum allowed depth of the tree
RandomForest	mtry (1:len(predictors); 1)	Number of randomly selected predictors at the node
	ntree (100, 250, 500, 750, 1000)	Number of trees in an ensemble
GBM	n.trees (150, 200, 250, 500)	Number of trees to learn
	interaction.depth (5:18; 1)	Number of splits per tree
	shrinkage (0.075, 0.1, 0.125, 0.15, 0.2)	Learning rate of the gradient
	n.minobsinnode (3, 5, 7, 10, 12, 15)	Minimum number of observations per node
XGBoost	eta (0.01, 0.1, 0.2, 0.3)	Learning rate of the gradient
	max_depth (4:8; 1)	Maximum depth of a tree
	min_child_weight (1, 3, 5, 7)	Minimum sum of instance weight needed in a child
	subsample (0.65, 0.8, 1)	Subsample ratio of the training instances
	colsample_bytree (0.8, 0.9, 1)	Subsample ratio of columns when constructing each tree
	nrounds (1000)	Number of rounds for boosting
GLMNet	alpha (0:1; 1)	Ratio of mixing Ridge (alpha=0) and Lasso (alpha=1) Regression
	lambda (0.0001:1; 0.01)	Strength of the influence of the regression type

C.2 Algorithm to model training with multiple oversampling rates

Algorithm 1: Algorithm for training models with multiple oversampling rates.

```
Data: data ← merge(taskData, subjectData)
avgResults ← data.frame()
depVars ← {sls, slrt}
models ← {DummyRegressor, rPart, RandomForest, GBM, GLMNet}
foreach depVar in depVars do
    i ← 0
    while i < 5 do
        iResults ← data.frame()
        trainData ← stratifiedPartition(data * 0.7)
        oversamplingRates ← {0, 1, 1.5, 2, 2.5, 3, 3.5, 4}
        foreach oRate in oversamplingRates do
            if rate != 0 then
                trainData ← oversample(trainData, oRate)
            end
            foreach model in models do
                model ← train(model, trainData)
                iResults_oRate += rmse_model
            end
        end
        i++
    end
    foreach model in models do
        foreach oRate in oversamplingRates do
            avgResults_oRate = mean(iResults_model rmse)
        end
    end
end
```

C.3 Model training results

Table C.2: We trained each model on a data set, which was oversampled from 0% to 400%. The RMSE indicates how well each model performed during model training. The percentage change indicates how oversampling influenced the RMSE.

		Reading		Spelling	
Model	OS rate	RMSE	change in %	RMSE	change in %
Dummy Regressor	0.00	8.81	100.00	5.10	100.00
Dummy Regressor	1.00	9.08	103.07	5.23	102.63
Dummy Regressor	1.50	9.28	105.28	5.37	105.20
Dummy Regressor	2.00	9.50	107.81	5.51	108.03
Dummy Regressor	2.50	9.73	110.41	5.66	111.00
Dummy Regressor	3.00	9.95	112.90	5.80	113.78
Dummy Regressor	3.50	10.18	115.55	5.95	116.73
Dummy Regressor	4.00	10.40	117.99	6.09	119.45
GBM	0.00	9.06	100.00	5.20	100.00
GBM	1.00	9.14	100.90	5.36	103.11
GBM	1.50	9.26	102.18	5.30	102.01
GBM	2.00	9.30	102.62	5.45	104.83
GBM	2.50	9.53	105.20	5.40	103.98
GBM	3.00	9.23	101.89	5.42	104.21
GBM	3.50	9.22	101.72	5.39	103.74
GBM	4.00	9.22	101.73	5.38	103.60
GLMNet	0.00	8.64	100.00	5.00	100.00
GLMNet	1.00	8.65	100.05	5.13	102.70
GLMNet	1.50	8.84	102.27	5.19	103.98
GLMNet	2.00	9.00	104.07	5.28	105.80
GLMNet	2.50	9.08	105.06	5.38	107.75
GLMNet	3.00	9.21	106.59	5.45	109.09
GLMNet	3.50	9.36	108.27	5.53	110.64
GLMNet	4.00	9.49	109.78	5.61	112.32
Random Forest	0.00	8.53	100.00	4.89	100.00
Random Forest	1.00	8.56	100.39	4.95	101.25
Random Forest	1.50	8.56	100.34	4.96	101.57
Random Forest	2.00	8.52	99.93	4.98	101.94
Random Forest	2.50	8.58	100.63	4.99	102.14
Random Forest	3.00	8.51	99.84	5.00	102.36
Random Forest	3.50	8.56	100.41	5.01	102.46
Random Forest	4.00	8.56	100.36	5.02	102.72
rPart	0.00	8.60	100.00	5.12	100.00
rPart	1.00	8.95	104.11	5.35	104.42
rPart	1.50	8.97	104.36	5.43	106.08
rPart	2.00	9.12	106.13	5.50	107.32
rPart	2.50	9.26	107.68	5.74	112.17
rPart	3.00	9.34	108.69	5.86	114.42
rPart	3.50	9.76	113.56	5.94	116.03
rPart	4.00	9.77	113.66	6.05	118.11