

# Early Screening of Dyslexia Using a Language-Independent Content Game and Machine Learning

## Maria Rauschenberger

---

DOCTORAL THESIS UPF / 2019

Directors of the thesis:

Prof. Dr. Ricardo Baeza-Yates  
Department of Information and Communication Technologies,  
Universitat Pompeu Fabra

Dr. Luz Rello  
Department of Information Systems and Technology,  
IE Business School, IE University



*Dedicated to people with dyslexia*

*“For your own sanity, you have to remember that not all problems can be solved. Not all problems can be solved, but all problems can be illuminated.” – Ursula Franklin*



# Acknowledgements

---

*thanks to...*

All the wonderful people I met along the way. Some of them I knew before I decided to do a Ph.D. and some I got to know during the *Ph.D. Journey*. All of you are part of my journey, and I am very happy to be part of yours:

talking about methods, results or conferences;  
doing trips, hikes, or holidays;  
going for lunch, volleyball, a *vermouth*, dancing, or beers;  
or just hanging out;  
I want to thank all of you!!!! Please feel appreciated!

Special thanks to my supervisors Luz and Ricardo, with whom I had a wonderful research experiences and who gave me freedom and trust throughout my work! Thank you for your steady leadership and for sharing your expertise when I needed it. Both have made me a more mature researcher and person.

A Ph.D. is not something you can do alone. For me, it feels like a community project in which a network of people is working on a goal, and I am connecting the dots. I hope that my work can serve the people who supported me and this project! Thank you to every child, mother, father, teacher, headmaster, therapist, and supporter who participated in my study and helped me to spread the word of my research! Here, I would like to first name the three people who supported me through the years: Friederike Hansch, Dr. Monika Batke, and Prof. Dr. Jörg Thomaschewski. Thank you for your support! Second, thank you to Yeliz Yesilada and Jeffrey Bigham for your report to achieve the *international mention*. Third,

thank you to the committee members for my Ph.D. defense: Sergi Grau, Simon Harper, and Patricia Santos.

Ein herzliches Dankeschön an meine Familie — Mama, Papa, Schwestern, Schwägern, Nichten und Neffen: Ihr macht es mir sehr einfach über andere Dinge nachzudenken! Dank auch an dich mein Schatz, denn deine Ruhe, Geduld und Unterstützung sowie die gemeinsame Zeit sind einfach wunderbar!

Sometimes I have the feeling that dealing with dyslexia in my childhood prepared me very well for a life in academia and computer science. Both require me to have a high level of tolerance for errors and to find new successful solutions. Therefore, I would like to encourage everyone to follow what they like and to search for their strengths! I would especially like to encourage children with dyslexia and their parents to believe they can achieve great things.

Thanks to all collaborators, colleagues, friends and staff in Barcelona, Emden, Oldenburg, and Pittsburgh. Staying and collaborating with different labs, people, approaches, and cultures have enriched my research and me as a person.

This thesis was partially supported by an ICT Ph.D. program scholarship of Universitat Pompeu Fabra and by the *fem:talent Scholarship* from the *Applied University of Emden/Leer* as well as by the *Deutschen Lesepreis 2017* from the *Stiftung Lesen* and the *Commerzbank-Stiftung*.

Thank you all!

P.S. Since there are so many people to thank, I have added a second part of the Acknowledgment in the Appendix.

# Abstract

---

Children with dyslexia have difficulties learning how to read and write. They are often diagnosed after they fail in school, even though dyslexia is not related to general intelligence. In this thesis, we present an approach for earlier screening of dyslexia using a language-independent game in combination with machine learning models trained with the interaction data. By earlier, we mean before children learn how to read and write.

To reach this goal, we designed the game content with knowledge of the analysis of word errors from people with dyslexia in different languages and the parameters reported to be related to dyslexia, such as auditory and visual perception. With our two designed games (MusVis and DGames), we collected data sets (313 and 137 participants) in different languages (mainly Spanish and German) and evaluated them with machine learning classifiers. For MusVis we mainly use content that refers to one single acoustic or visual indicator, while DGames content refers to generic content related to various indicators. Our method provides an accuracy of 0.74 for German and 0.69 for Spanish and F1-scores of 0.75 for German and 0.75 for Spanish in MusVis when Random Forest and Extra Trees are used. DGames was mainly evaluated with German and reached a peak accuracy of 0.67 and a peak F1-score of 0.74. Our results open the possibility of low-cost and early screening of dyslexia through the Web.

# Resum

---

Els nens amb dislèxia tenen dificultats per aprendre a llegir i escriure. Sovint se'ls diagnostica després de fallar a l'escola, encara que la dislèxia no estigui relacionada amb la intel·ligència general. En aquesta tesi, presentem un enfocament per a la selecció prèvia de la dislèxia mitjançant un joc independent del llenguatge en combinació amb models d'aprenentatge automàtic formats amb les dades d'interacció. Abans volem dir abans que els nens aprenguin a llegir i escriure.

Per assolir aquest objectiu, vam dissenyar el contingut del joc amb el coneixement de l'anàlisi de paraules d'error de persones amb dislèxia en diferents idiomes i els paràmetres relacionats amb la dislèxia com la percepció auditiva i la percepció visual. Amb els nostres dos jocs dissenyats (MusVis i DGames) vam recollir conjunts de dades (313 i 137 participants) en diferents idiomes (principalment espanyols i alemanys) i els vam avaluar amb classificadors d'aprenentatge automàtic. Per a MusVis utilitzem principalment contingut que fa referència a un únic indicador acústic o visual, mentre que el contingut de DGames fa referència a diversos indicadors (també contingut genèric). El nostre mètode proporciona una precisió de 0,74 per a l'alemany i 0,69 per a espanyol i una puntuació de F1 de 0,75 per a alemany i de 0,75 per a espanyol a MusVis quan s'utilitzen arbres extraestats. DGames es va avaluar principalment amb alemany i obté la màxima precisió de 0,67 i la màxima puntuació de F1 de 0,74. Els nostres resultats obren la possibilitat de la dislèxia de detecció precoç a baixos costos a través del web.

# Resumen

---

Los niños con dislexia tienen dificultades para aprender a leer y escribir. A menudo se les diagnostica después de fracasar en la escuela, incluso aunque la dislexia no está relacionada con la inteligencia general. En esta tesis, presentamos un enfoque para la detección temprana de la dislexia utilizando un juego independiente del idioma en combinación con modelos de aprendizaje automático entrenados con los datos de la interacción. Temprana aquí significa antes que los niños aprenden a leer y escribir.

Para alcanzar este objetivo, diseñamos el contenido del juego con el conocimiento del análisis de las palabras de error de las personas con dislexia en diferentes idiomas y los parámetros reportados relacionados con la dislexia, tales como la percepción auditiva y la percepción visual. Con nuestros dos juegos diseñados (MusVis y DGames) recogimos conjuntos de datos (313 y 137 participantes) en diferentes idiomas (principalmente español y alemán) y los evaluamos con clasificadores de aprendizaje automático. Para MusVis utilizamos principalmente contenido que se refiere a un único indicador acústico o visual, mientras que el contenido de DGames se refiere a varios indicadores (también contenido genérico). Nuestro método proporciona una exactitud de 0,74 para alemán y 0,69 para español más una puntuación F1 de 0,75 para alemán y 0,75 para español en MusVis cuando se utilizan Random Forest y Extra Trees, respectivamente. DGames fue evaluado principalmente con alemán, obteniendo una exactitud de 0,67 y una puntuación F1 de 0,74. Nuestros resultados abren la posibilidad de una detección precoz y de bajo coste de dislexia a través de la Web.



# Abstrakt

---

Kinder mit einer Lese-/Rechtschreibstörung (LRS) haben Schwierigkeiten, Lesen und Schreiben zu lernen. Sie werden oft nach dem Schulversagen diagnostiziert, auch wenn die LRS unabhängig von der allgemeinen Intelligenz ist. In dieser Arbeit stellen wir einen Ansatz für ein früheres Screening von der LRS mit einem sprachunabhängigen Spiel in Kombination mit machinellen Lernmodellen vor, die mit den Interaktionsdaten trainiert wurden. Mit früher meinen wir, bevor Kinder das Lesen und Schreiben lernen.

Um dieses Ziel zu erreichen, haben wir Spielinhalte mit dem folgenden Wissen erstellt: die Analyse von Fehlerwörtern von Menschen mit einer (LRS) in verschiedenen Sprachen und die Parameter, die mit der LRS zusammenhängen, wie z.B. auditive Wahrnehmung und visuelle Wahrnehmung. Mit unseren beiden entwickelten Spielen (MusVis und DGames) haben wir Datensätze (313 und 137 Teilnehmer/innen) in verschiedenen Sprachen (hauptsächlich Spanisch und Deutsch) gesammelt und mit *Machine Learning Classifier* ausgewertet. Wir verwenden für MusVis hauptsächlich Inhalte, die sich auf einen einzigen akustischen oder visuellen Indikator beziehen, während DGames-Inhalte auf verschiedene Indikatoren (auch generische) bezogen sind. Unsere Methode liefert für MusVis eine Genauigkeit von 0,74 für Deutsch und 0,69 für Spanisch und einen F1-Score von 0,75 für Deutsch und 0,75 für Spanisch in MusVis, wenn Random Forest und Extra Trees verwendet werden. DGames wurde hauptsächlich mit deutscher Sprache ausgewertet und erreicht die höchste Genauigkeit von 0,67 und den höchsten F1-Score von 0,74. Unsere Ergebnisse eröffnen die Möglichkeit, die Lese-/Rechtschreibstörung kostengünstig und frühzeitig über das Internet zu erkennen.

# Contents

---

<b>Abstract</b>	<b>vii</b>
<b>Resum</b>	<b>viii</b>
<b>Resumen</b>	<b>ix</b>
<b>Abstrakt</b>	<b>x</b>
<b>List of Figures</b>	<b>xviii</b>
<b>List of Tables</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Goals . . . . .	6
1.3 Challenges . . . . .	7
1.4 Contributions . . . . .	8
1.5 Structure . . . . .	9
<b>2 Background</b>	<b>11</b>
2.1 Introduction . . . . .	11

2.2	Dyslexia . . . . .	11
2.2.1	Dyslexia Screening . . . . .	15
2.2.2	Screening for Readers . . . . .	16
2.2.3	Screening for Pre-Readers . . . . .	23
2.2.4	Auditory Perception . . . . .	30
2.2.5	Visual Perception . . . . .	32
2.3	Design . . . . .	33
2.3.1	Design Science Research Methodology . . . . .	33
2.3.2	Human-Centred Design . . . . .	34
2.3.3	General Considerations of Research Design . . . . .	38
2.4	Gamification . . . . .	41
2.4.1	The Evolution of Gamification . . . . .	42
2.4.2	Gamification vs. Serious Games . . . . .	43
2.5	Summary . . . . .	45
<b>3</b>	<b>Methodology</b> . . . . .	<b>47</b>
3.1	Introduction . . . . .	47
3.2	Combining HCD and DSRM . . . . .	47
3.3	Problem Identification . . . . .	51
3.4	Concept . . . . .	52
3.4.1	Objectives . . . . .	53
3.4.2	Participant Requirements and Context . . . . .	54
3.4.3	Integrating Gamification . . . . .	55
3.5	Content Design . . . . .	56
3.6	Demonstration . . . . .	56
3.6.1	Experimental Design . . . . .	57
3.6.2	Ethics . . . . .	58
3.6.3	Data Collection . . . . .	59
3.7	Evaluation . . . . .	60
3.8	Outreach . . . . .	63
3.9	Discussion . . . . .	64
<b>4</b>	<b>Screening Dyslexia with Auditory and Visual Cues</b> . . . . .	<b>67</b>
4.1	Introduction . . . . .	67

4.2	Methodology . . . . .	69
4.3	Game Design . . . . .	69
4.3.1	Auditory Cues . . . . .	72
4.3.2	Visual Cues . . . . .	77
4.3.3	User Interface . . . . .	80
4.3.4	Implementation . . . . .	80
4.4	Usability Test . . . . .	81
4.4.1	Procedure . . . . .	81
4.4.2	Participants . . . . .	82
4.4.3	Usability Improvements . . . . .	82
4.5	Experimental Design Setup . . . . .	84
4.5.1	Procedure . . . . .	84
4.5.2	Participant Groups . . . . .	86
4.5.3	Dependent Variables and Features . . . . .	89
4.6	Predictive Models Setup . . . . .	93
4.6.1	Model Selection . . . . .	93
4.6.2	Feature Selection . . . . .	93
4.7	Statistical Analysis . . . . .	94
4.7.1	Pilot-Study . . . . .	95
4.7.2	Validation . . . . .	101
4.8	Prediction using Machine Learning . . . . .	103
4.9	Discussion . . . . .	106
4.9.1	Group Comparison . . . . .	107
4.9.2	Screening Differences . . . . .	109
<b>5</b>	<b>Screening Dyslexia adding Generic Content</b>	<b>113</b>
5.1	Introduction . . . . .	113
5.2	Methodology . . . . .	114
5.3	Game Design . . . . .	115
5.3.1	Auditory Cues adding Generic Content . . . . .	117
5.3.2	Visual Cues adding Generic Content . . . . .	119
5.3.3	User Interface . . . . .	121
5.3.4	Implementation . . . . .	122
5.4	Experimental Design Setup . . . . .	123

5.4.1	Procedure . . . . .	123
5.4.2	Participants . . . . .	123
5.4.3	Dependent Variables and Features . . . . .	126
5.5	Predictive Models Setup . . . . .	131
5.5.1	Model Selection . . . . .	131
5.5.2	Feature Selection . . . . .	132
5.6	Results . . . . .	134
5.6.1	Statistical Analysis . . . . .	134
5.6.2	Prediction using Machine Learning . . . . .	138
5.7	Discussion . . . . .	141
5.7.1	Statistical Comparison . . . . .	141
5.7.2	Screening Differences . . . . .	143
<b>6</b>	<b>Conclusions and Future Work</b>	<b>145</b>
6.1	Summary . . . . .	145
6.2	Future Directions . . . . .	150
	<b>Bibliography</b>	<b>152</b>
<b>A</b>	<b>Appendix</b>	<b>177</b>
A.1	Towards the Use of Gamification . . . . .	177
A.2	Analysis of German Error Words . . . . .	182
A.3	Study Approvals . . . . .	185
A.4	Further Acknowledgements . . . . .	185

# List of Figures

---

1.1	Content structure of the thesis. . . . .	10
2.1	Example exercises from the dyslexia screener <i>Dy-tective English</i> : <b>a)</b> player needs to click the tar-get non-word listed among the distractors; <b>b)</b> player needs to click on the different letter [129]. . . . .	18
2.2	Screen examples: <b>a)</b> <i>Paths</i> game [43]; <b>b)</b> <i>Lexa</i> [98]; <b>c)</b> <i>Fence letters game</i> [43]; and <b>d)</b> <i>DYSL-X</i> [152]. . . . .	28
2.3	Activities of the human-centred design process visu-alised by the author from [68]. . . . .	36
2.4	Approach of cross-validation from [141]. . . . .	40
2.5	Our synthesis from elements of game design to dif-ferent outcomes. . . . .	44
3.1	Integration of the <i>Human-centred Design</i> in the <i>De-sign Science Research Methodology</i> . . . . .	48
3.2	Overview of our actions for the human-centred design. . . . .	50
4.1	Participants playing the visual part (left) and the au-ditory part (right) of the Game <i>MusVis</i> . Photos in-cluded with permission. . . . .	70

4.2	Example of the auditory part from the game <i>MusVis</i> for the first two clicks on two sound cards (left) and then a pair of equal sounds is found (right). The participant is asked to find two equal auditory cues by clicking on sound cards. . . . .	71
4.3	Example of the visual part of the game <i>MusVis</i> with the priming of the target cue <i>symbol</i> (left) and the nine-squared design including the distractors for each <i>symbol</i> (right). . . . .	72
4.4	Waveform for the order of intervals for one auditory cue of the stage <i>Rise Time</i> . The example starts with a 0.025s fade in interval and then a 0.250s fade in interval followed by a 0.250s fade in interval. . . . .	77
4.5	Overview of the designed visual cues. The figure shows the target cue (top) and distractor cues (below) for the four different stages ( <i>z</i> , <i>symbol</i> , <i>rectangle</i> , <i>face</i> ) of the visual part of the game <i>MusVis</i> . . . . .	79
4.6	Normalised confusion matrix from the three best results (F1-score and accuracy): <b>a) DE, 5 features with RF</b> ; <b>b) ES, 20 features with ETC</b> ; and <b>c) ALL, 20 features with GB</b> . . . . .	105
4.7	The plot shows the relation of accuracy to features for all classifiers in the data set ALL (left), ES (middle) and DE (right). . . . .	106
5.1	Example of the auditory part of the game <i>DGames</i> with the priming of the target cue (a) and then the distractors for each auditory cue (b). . . . .	120
5.2	Example of the visual part of the game <i>DGames</i> with the priming of the target cue <i>animal</i> (a) and then the four-squared (b) and (c) nine-squared design including the distractors for each <i>animal</i> . . . . .	121

5.3	Overview of the designed related-linguistic (called <i>related</i> ) and non-linguistic (called <i>generic</i> ) cues. The figure shows the target cue (top) and distractor cues (below) for the eight different stages ( <i>symbol, z, rectangle, face, fruit, kitchen, plant, animal</i> ) of the visual part of the game <i>DGames</i> . . . . .	122
5.4	The plot shows the features ranking for the DE and ALL data sets with the highest-ranked features highlighted. Feature ID follows the index described in Section 5.4.3. . . . .	132
5.5	Normalised confusion matrix from the two best results (F1-score and accuracy): <b>a)</b> <i>ALL, Informativ with ETC</i> ; <b>b)</b> <i>DE, Informativ with ETC</i> ; <b>c)</b> <i>ALL, Auditory generic with ETC</i> ; and <b>d)</b> <i>DE, Auditory generic with ETC</i> . . . . .	140
A.1	Overview of the abstraction levels: 115 Dynamics, Mechanics and Components. . . . .	179
A.2	Distribution of dynamics ( $n = 115$ ). . . . .	180
A.3	Distribution of the error position for the German error resource [109] for position 0 to 12. . . . .	182
A.4	Distribution of the error categories for the German error resource [109]. . . . .	183
A.5	MusVis study approval by the Ministry of Education, Science and Culture of Schleswig-Holstein ( <i>Ministerium für Bildung, Wissenschaft und Kultur, MBWK</i> ) in German. . . . .	187
A.6	MusVis study approval by the <i>Education Authority</i> of the State of Lower Saxony ( <i>Niedersächsische Landesschulbehörde</i> ) in German (part one). . . . .	188
A.7	MusVis study approval by the <i>Education Authority</i> of the State of Lower Saxony ( <i>Niedersächsische Landesschulbehörde</i> ) in German (part two). . . . .	189



A.8 DGames study approval by the Ministry of Education, Science and Culture of Schleswig-Holstein (*Ministerium für Bildung, Wissenschaft und Kultur, MBWK*) in German. . . . . 190

A.9 DGames study approval by the *Education Authority* of the State of Lower Saxony (*Niedersächsische Landesschulbehörde*) in German (part one). . . . . 191

A.10 DGames study approval by the *Education Authority* of the State of Lower Saxony (*Niedersächsische Landesschulbehörde*) in German (part two). . . . . 192

# List of Tables

---

2.1	Cognitive skills tested in different dyslexia screening tools for readers (part one). . . . .	19
2.2	Cognitive skills tested in different dyslexia screening tools for readers (part two). . . . .	20
2.3	Evaluation of reader screening tools. . . . .	21
2.4	Cognitive skills tested in dyslexia screeners for pre-readers. . . . .	25
2.5	Study details pre-reader screening tools if results are published. . . . .	26
3.1	Overview of the data-driven approach. . . . .	57
4.1	Auditory cues generated for the four tasks in <i>MusVis</i> . * 2 cues: (1/2 semitone - 50 cents interval); $\Delta$ 3 cues: 3 sounds spaced by 25 cents (quarter of a semitone) - 2 previous ones. . . . .	74
4.2	Description of the auditory attributes which show promising relations to the prediction of dyslexia. . . . .	75

4.3	Mapping of the evidence from literature to distinguish a person with dyslexia, the attributes and general assumptions, and the stages of the auditory part of the game <i>MusVis</i> . . . . .	76
4.4	Overview of the participants per data set for the validation experiments. . . . .	88
4.5	Description of participant features. . . . .	91
4.6	On the left are features 10 to 105 for the auditory part and on the right are features 106 to 201 for the visual part of the game <i>MusVis</i> . . . . .	92
4.7	Overview of all selected dependent variables for the auditory and visual parts of the game <i>MusVis</i> for German. . . . .	96
4.8	Overview of all selected dependent variables for the auditory and visual part of the game <i>MusVis</i> for Spanish. . . . .	97
4.9	Overview of dependent variables with the same tendency, which are all from the visual part of the game for ALL. . . . .	97
4.10	Overview of all dependent variables showing the language-independent results between the German and Spanish groups. . . . .	98
4.11	Overview of dependent variables for visual (top) and auditory (below) features of <i>DGames</i> . Significant results are in bold. . . . .	102
4.12	Best results of the different classifiers, features and data sets. Results are ordered by the best F1-score and accuracy. . . . .	104
5.1	Description of the auditory attributes for <i>DGames</i> . . . . .	118
5.2	Mapping of the evidence from literature to distinguish a person of dyslexia to design the auditory type for each stage of <i>DGames</i> . . . . .	119
5.3	Overview of the participants per data set. . . . .	124

5.4	Overview of bilingualism per data set. *One Language; $\Delta$ Bilingualism. . . . .	125
5.5	Overview of the dependent variables used for the statistical comparison. . . . .	126
5.6	Description of the participant features. . . . .	128
5.7	Description of the auditory features. . . . .	129
5.8	Description of the visual features (part one). . . . .	130
5.9	Description of the visual features (part two). . . . .	131
5.10	DE and ALL have 15 features in common among the highest-ranked informative features (DE $n = 53$ , ALL $n = 58$ ). . . . .	133
5.11	Overview of the subsets of features used to compare the quality of the prediction. . . . .	134
5.12	Overview of all reported dependent variables for the auditory (top) and visual (below) part of the game <i>DGames</i> for DE ( $n = 120$ ). Significant results are in bold. . . . .	136
5.13	Best results of ALL (on the left) and DE (on the right) data sets for the different classifiers and subsets of features. The best two results for the F1-score and accuracy are highlighted as well as difference in the classifier ranking. . . . .	139
6.1	Cognitive skills tested in dyslexia screeners for pre-readers. * <i>DGames</i> addresses the same skills as <i>MusVis</i> . . . . .	149
A.1	Distribution of the Damerau-Levenshtein distance [26, 78] for the German error words. . . . .	184



# Introduction

---

## 1.1 Motivation

In today's world, not being able to read and write properly is a huge disadvantage in society. Despite their normal intelligence, individuals with dyslexia have difficulties learning reading and writing. Dyslexia is a *specific learning disorder* which affects 5% to 15% of the global population [2].

Dyslexia in children is often detected by spelling and reading mistakes, as well as a lack of reading and writing skills that indicates academic failure. A person with dyslexia demonstrates normal or high levels of intellectual functioning [2] and is thus able to consciously compensate for these deficits [87], making dyslexia hard to detect. Observable manifestations of dyslexia typically emerge when a child reaches a certain age and literacy level. This is why current approaches for screening pre-readers require expensive personnel (such as a professional therapist) or special hardware (such as MRI or fMRI scanners [94, 95, 150] or eye-tracking [4]). Also, most of the methods can only be used when children are learning how to read but not before, which delays needed early intervention.

Detecting dyslexia is important because early intervention is key to avoiding the negative effects of it, such as school failure or negative thoughts [138]. The literature shows that dyslexia is related to various non-linguistic indicators such as short-term memory [47], visual-spatial attention [38], motor skills [87], *reduced phonological information processing* [103], auditory perception [47], or phonological working memory [103]. Furthermore, Nicolson and Fawcett provide evidence that children with dyslexia show *lapses of concentration* when applying cognitive or motor skills [87]. To sum up, prior work found evidence in lab studies of language-independent indicators associated with dyslexia, such as visual-spatial attention and phonological working memory.

Our work shows how to design language-independent content with non-linguistic indicators for universal dyslexia screening using games and a machine learning prediction using the interaction data. At this point, a long-term study with pre-readers would be very time-consuming, since the effort to find participants is high, participants are less likely to be diagnosed, and much time passes before results are available. An online study with readers has the advantage of reducing the effort and time required to design content, conduct various experiments for optimization, and increase the number of participants. Nevertheless, the language-independent content can be used to screen pre-readers who do not yet have any language skills.

We consider our research game not as a medical detection tool but as an objective, uncomplicated, user-friendly web screening tool. Such a web game has the potential to be low-cost and easily accessible, as it is available for a broader audience through the Web. It can also serve to make parents aware of the risks of dyslexia and guide them to additional help and resources (e.g., medical doctor or therapist).

Given the previous context, there are three main motivations behind our work, each one with a high social impact. First, dyslexia occurs frequently and across different languages and cultures [2, 104,

163, 164] and can have strong negative effects on individuals [138]. Second, a language-independent screening tool could be the starting point of early detection for pre-readers, which is a challenge [6]. Hence, children have more time to learn compensation strategies, which reduces the pressure placed on them and increases their chances for success. Third, a low-cost, playful, and easily adaptable screening method can reach a wide population, making people aware of the risk of dyslexia as well as guiding parents to more help.

A person with dyslexia has difficulties with reading and writing that are independent from intelligence, mother tongue, social status, or education level. Hence, people with dyslexia understand the meanings of words, but do not always know how to spell or pronounce them correctly. However, children with dyslexia do not show any obvious difficulties in other areas. This is why *dyslexia* is considered to be a *hidden* disorder. This often results in bad grades in school and frustration for the children and parents over many years. Around 40% to 60% of children with dyslexia show symptoms of psychological disorders [138] such as negative thoughts, sadness, sorrow, or anxiety. A study showed that even if the child is diagnosed by the age of eight, they achieve lower school performance [31]. Also, according to the same study, the unemployment rate for adults with dyslexia is higher. Moreover, these are common indicators for detecting a person with dyslexia.

The *International Dyslexia Association* states that “15% to 20% of the population has a language-based learning disability” [65]. Even though language acquisition depends on the syllabic complexity and orthographic depth of a language [145], results show that similarities between readers with dyslexia in English and in German are far bigger than their differences [167]. Also, similar types of errors were found in texts written by people with dyslexia for English, Spanish [123], and German [118]. Multiple factors have been investigated to discover the causes of dyslexia, how to measure it, and which skills need to be trained to improve reading and writing [21].



Current approaches such as the German test *Diagnostische Rechtschreibtest* [51] detect dyslexia mainly by the spelling and reading mistakes of a child, which requires the child to have linguistic skills [2, 23, 140]. The reason early prediction of dyslexia is considered challenging [6] is that dyslexia manifests itself in reading and writing, skills which have not yet been acquired by pre-readers. However, early detection is needed, since it is possible for a person with dyslexia to gain reading comprehension and spelling accuracy with appropriate intervention. As a matter of fact, children with dyslexia can learn the spellings of words or decode words for reading, but they need more time to practice [138]. For example, children need two years instead of one for learning how to spell phonetically accurate words. Hence, to provide children with dyslexia more time to practice, help them to avoid frustration, and give them the possibility to succeed, early, language-independent detection is needed. Our overall hypothesis is therefore:

**Dyslexia is normally detected using linguistic elements. That is only possible when children have already developed reading skills. Is it possible to detect a child at risk for dyslexia with language-independent content?**

Language-independent content could be used to screen pre-readers for dyslexia as well as to screen dyslexia across languages. Dyslexia has been studied extensively, but no scientific agreement on the causal origin has been achieved [15]. Currently, there are two main theories, considering either visual [156] or auditory [46] perception to be a critical causal component. It has been argued that dyslexia might be mainly rooted in phonological and perception differences [46]. Non-digital approaches, e.g., [70], try to predict the literacy skills of children with phonological perception, phonological working memory processing, long-term memory, and visual attention (quoted after [148]). Another line of research suggests that reading difficulties are due to problems with

visual-spatial attention and poor coding rather than phonological difficulties [156]. In fact, non-similar sounds might be used as a compensation strategy to cope with dyslexia, which breaks down when we have phonetic ambiguity. In other words, we see a symptom of the problem, but not the real cause. To conclude, the literature shows evidence of non-linguistic indicators related to dyslexia. These indicators could be integrated into a game designed to screen for dyslexia, which could also address the challenge of early dyslexia screening [6].

Lately, it has been shown that computer games are a convenient medium to easily, quickly, and cost-efficiently screen children with dyslexia using a web tool that, among other things, analyzes word errors from people with dyslexia [127]. However, collecting health data is costly in terms of time and resources [34]. Recently, machine learning has been used to predict dyslexia, using people with and without dyslexia to distinguish differences between small groups [6]. However, when machine learning techniques are applied to small data, precautions of over-fitting need to be addressed [30].

This prior work motivated us to design our language-independent game content with auditory and visual cues to analyze the differences in game measures between children with and without dyslexia. For the content design of the auditory and visual cues, prior language acquisition, phonological awareness, letter naming, and letter recognition are not required. Combining the language-independent content with a web game can make the screening playful and easily accessible for a wider population. An additional advantage of a language-independent approach is that only the instructions for the supervisor need to be translated. Such a low-cost, playful, and easy-to-use screening tool could raise awareness of dyslexia, guide parents to more help, and therefore support 5% to 15% of the child population.

## 1.2 Goals

In our work, we first aim to provide predictive results on the universal screening of dyslexia with language-independent content using games and a machine learning prediction using the interaction data. There are three secondary goals:

- To design language-independent content that can be used as input in a game to measure differences between children with and without dyslexia.
- To design web applications as a low-cost approach for non-professionals to conduct a quick and preliminary screening of people who may have dyslexia and should see a professional.
- To find significant dependent measures to distinguish children with and without dyslexia, as well as to screen children for dyslexia using machine learning classifiers.

We aim to reach our goals for this thesis by answering the following research questions:

- R1** Are there significant statistical differences between children with and without dyslexia when playing a game with auditory and visual content?
- R2** Is it possible to predict risk of dyslexia based on language-independent auditory and visual content using a game and machine learning for different languages?
- R3** Is it possible to predict risk of dyslexia based on generic language-independent visual and auditory content with various indicators using a game and machine learning?

## 1.3 Challenges

Early, accurate prediction of dyslexia remains a challenge [6] because dyslexia is known for causing reading and writing problems but no obvious deficits in other areas. As explained in the motivation section, this is why dyslexia is referred to as a *hidden* disorder. Therefore, we need to design language-independent content fit to differentiate between children with and without dyslexia.

Another challenge is finding language-independent content that can show measurable differences between children with and without dyslexia that are comparable to differences in reading and writing mistakes. Designing language-independent content is probably the greatest challenge (also according to a report from the *National Center on Improving Literacy* [97]) because the new indicators, though related to the reading and writing difficulties, are probably not the main causes. Additionally, our baseline for separating the participant groups is affected by the different standards of diagnostic tools and the high variance of dyslexia. Furthermore, this new content needs to be integrated into a game context, as well as be designed to be used in an online experiment as a game. Previously studied language-independent indicators have been used in lab settings, which means these indicators have been tested in controlled environments. That is not the case for online experiments. Consequently, external factors must be controlled and influences made transparent for the analysis. We need to design the game with the proper indicators (content) and game constraints in order to collect dependent measures that reveal differences between the participant groups.

A further challenge for this thesis is in using small data to predict dyslexia with existing machine learning and without over-fitting. This will be a challenge throughout the thesis, as engaging participants over the Web, specifically parents of children 7 to 12 years old, requires a lot of time, communication, and recruitment.

## 1.4 Contributions

The main contributions of this Ph.D. are the game content design for the experiments, the data sets collected, and the first screening results with language-independent content using a game and machine learning prediction techniques. These games provide first results on screening a person at risk for dyslexia with language-independent content using a machine learning prediction using the interaction data. They may also give children time to practice skills, help them to gain confidence in their abilities, and increase their chances for success in school.

In the following, we list the publications that we published while working on this thesis.

- Rauschenberger, M., Willems, A., Ternieden, M., and Thomaschewski, J. (2019). Towards the use of gamification frameworks in learning environments. *Journal of Interactive Learning Research*, 30(2) (Section 2.4, Chapter 3 and Appendix A.1, [122]).
- Rauschenberger, M., Baeza-Yates, R., and Rello, L. (2019). Technologies for Dyslexia. In *Web Accessibility A Foundation for Research* (Second Edition). Springer-Verlag London <http://doi.org/10.1007/978-1-4471-7440-0> (Chapters 1, 2 and 6, [114]).
- Rauschenberger, M., Rello, L., and Baeza-Yates, R. (2018). 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*. Barcelona: ACM Press (Chapter 5, [113]).
- Rauschenberger, M., Rello, L., Baeza-Yates, R., and Bigham, J. P. (2018). Towards Language Independent Detection of Dyslexia with a Web-based Game. In *Proceedings of the*

*Internet of Accessible Things on - W4A '18*. Lyon, France: ACM Press <http://doi.org/10.1145/3192714.3192816> (Chapter 4; Section 2.2.4 and 2.2.5, [115]).

- 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 *Proceedings of the 14th Web for All Conference on The Future of Accessible Work - W4A '17* (pp. 1–4). Perth, Western Australia: ACM Press <http://doi.org/10.1145/3058555.3058565> (Chapter 4, [117]).
- Rauschenberger, M. (2016). DysMusic: Detecting Dyslexia by Web-based Games with Music Elements. In *The Web for All Conference Addressing Information Barriers – W4A'16*. Montreal, Canada: ACM Press (Chapter 4, [106]).

## 1.5 Structure

The thesis is organized as follows (overview in Figure 1.1). After the introduction (Chapter 1) is the background (Chapter 2), in which we provide a description of dyslexia, the state-of-the-art on screening for readers and pre-readers, the design approaches, and gamification. In Chapter 3 we explain the overall thesis methodology and the details of the experimental design. We show in Chapter 4 the design of the language-independent content and the significant measures. We also present the first prediction for screening the risk of dyslexia with language-independent content (auditory and visual) using a game and machine learning. In Chapter 5 we present the added generic content and content related to more characteristics of dyslexia, as well as the improved auditory gameplay. The thesis ends with our conclusion and proposed future lines of research.

»»» Chap.1 — Introduction



»»» Chap.2 — Background



»»» Chap.3 — Methodology

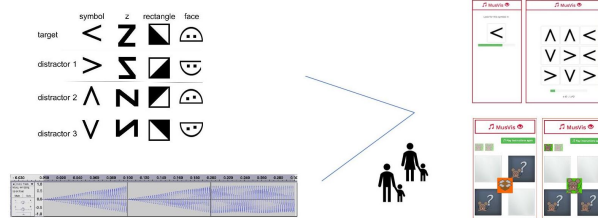


**R1:** Are there significant statistical differences between children with and without dyslexia when playing a game with auditory and visual content?

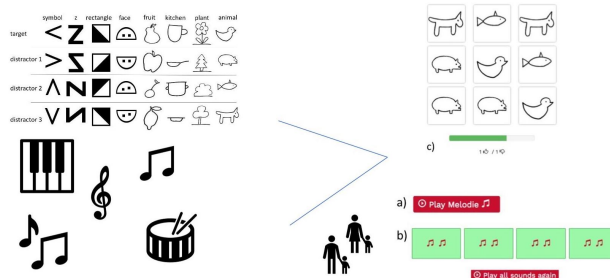
**R2:** Is it possible to predict risk of dyslexia based on language-independent auditory content using a game and machine learning for different languages?

**R3:** Is it possible to predict risk of dyslexia based on language-independent visual content using a game and machine learning for different languages?

»»» Chap.4 — Screening Dyslexia with Auditory and Visual Cues



»»» Chap.5 — Screening Dyslexia adding Generic Content



»»» Chap.6 — Conclusions and Future Work

Figure 1.1: Content structure of the thesis.

# Background

---

## 2.1 Introduction

In this chapter, we present the topics that are most relevant to our objectives into three parts: We start with an explanation of dyslexia and a review of the state-of-the-art that is relevant to our work. Then we present the core approaches used to design the thesis, the application, and the evaluation. We finish with an explanation of gamification. The content of Section 2.2 was published in [114, 115]. Parts of the content of Section 2.4 were published in [122].

## 2.2 Dyslexia

In this section, we explain the main characteristics of dyslexia and present the state-of-the-art related to the thesis, focusing on screening dyslexia for readers and pre-readers as well as on the visual and auditory perception of dyslexia.

The *American Psychiatric Organization* defines dyslexia as a *specific learning disorder* which affects around 5% to 15% of the global population [2]. A person with dyslexia has visual and audi-



tory difficulties with reading and writing independent of intelligence, native language, social status, or education level [163, 164]. The definition of dyslexia in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) is in harmony with the *International Statistical Classification of Diseases and Related Health Problems* (ICD – 10/ – 11) [2, 163, 164].

Hence, people with dyslexia understand the meanings of words but do not always know how to spell or pronounce the words correctly. Thus, children with dyslexia have no apparent difficulties in other areas, meaning that difficulties (reading and writing) can only be noticed after the children have gained literacy. This is why *dyslexia* is called a *hidden* disorder. Often, it results in bad grades in school and frustration for the children and the parents over several years. Between 40% and 60% of children with dyslexia show symptoms of psychological disorders [138] such as negative thoughts, sadness, sorrow, or anxiety. Moreover, these are common indicators for detecting a person with dyslexia.

Dyslexia occurs in different languages and cultures [2, 104], although language acquisition depends on the syllabic complexity and orthographic depth of a language [145]. Research shows that the similarities between readers with dyslexia in English and German are much greater than their differences [167]. Furthermore, errors made by persons with dyslexia are similar in English, Spanish [123] and German [118].

The treatment and diagnosis of dyslexia depends on the unique guidelines for each country, such as in Germany [29] or the United States of America [159]. Still, all guidelines are commonly based on the definitions of the DSM-5 [2] and/or ICD-10/-11 [163, 164].

As a matter of fact, children with dyslexia can learn the spelling of words or decode words for reading, but they need more time to practice. In Germany, for example, Schulte-Körne *et al.* state that children need two years instead of one for learning how to spell phonetically accurate words [138]. Although a person with dyslexia can improve reading comprehension and spelling accuracy, a certain

degree of difficulty will most likely remain, and assistive applications for reading and writing can be helpful. Overcoming dyslexia involves a great effort for children and requires doing language exercises regularly [60]. Schulte-Körne et al. [139] showed that it takes longer for children with dyslexia to achieve school grades that are equal to those of their peers, even if these children have an above-average socio-economic background. For example, in Germany, only 25% of the poor spellers achieve average spelling performance during the period of primary school [130]. Hence, in order to give children with dyslexia more time to practice and the possibility to succeed, early detection is needed.

Multiple factors have been investigated to discover the causes of dyslexia and measure it, as well as to understand which skills need to be trained to improve reading and writing [21, 27]. Currently, dyslexia is connected to nine genetic markers, and reading ability is highly hereditary [27]. However, dyslexia cannot be reduced to one cause; rather, it is a combination of factors. Therefore, the *DSM-V* has been updated to reflect the fact that “*most children with a specific learning disorder manifest deficits in more than one area*” [14]. For example, one genetic marker is connected to the “deficits in memory, phonological decoding, sight word reading, orthographic decoding, [...] and spelling” [27]. Another genetic marker is connected to *phonological awareness, spelling, phonemic decoding, and sight word reading* [27]. Yet, today’s dyslexia diagnoses (*i.e., spelling tests*) and/or treatment (*i.e., syllable training*) historically does not consider visual and auditory perception in addition to various indicators, but mainly addresses the spelling and reading mistakes [51, 148].

Over the last decades, dyslexia has been studied from different fields, but no scientific agreement of the causal origin has been achieved [15]. There are two main theories at this point [27]. One considers visual perception [156] to be a key attribute for the cause of dyslexia depending on the information processing and memory, while the other considers it to be auditory perception [46].

Researchers argue that dyslexia might be mainly based on phonological and perception differences [46]. Moreover, previous research has related speech perception difficulty to auditory processing, phonological awareness, and literacy skills [27, 133, 149]. Phonological deficits of dyslexia have also been linked to basic auditory processing [53]. The auditory perception of children with dyslexia has been proven to be related to sound structure [63] as well as to the auditory working memory [82]. Non-digital approaches, *e.g.*, [70], try to predict the literacy skills of children using phonological perception, phonological working memory processing, long-term memory, and visual attention (quoted after [148]). Another line of research suggests that reading difficulties are due to visual-spatial attention problems and poor coding instead of phonological difficulties [156]. Apart from this, researchers suggest visual discrimination and search efficiency as predictors for future reading acquisitions [38]. Recently, the missing visual asymmetry is proposed as one of the many possible causes of dyslexia [77].

Dyslexia is highly comorbid with other neuro-development disorders such as specific language impairment, attention-deficit hyperactivity disorder (ADHD) or dyscalculia [27]. For example, around 20% of children with dyslexia have ADHD [138]. As with dyslexia, ADHD [83] and dyscalculia [81] show difficulties for working memory. However, when compared, the different disorders show different deficits: *e.g.*, dyslexia corresponds with deficits in phonological perception, dyscalculia with deficits in visual perception, and ADHD with deficits in central executive functioning [81]. Maehler and Schuchardt [81] did not find interaction effects between the disorders and concluded that comorbidity causes additive working memory deficits.

Lately, it has been shown that computer games are a convenient medium for providing an engaging way to significantly improve the reading [44, 75] and spelling [44, 128] performances of children with dyslexia. Additionally, research showed that readers with dyslexia could be detected easily and cost-efficiently with a web tool that,

among other things, analyzes word errors from people with dyslexia [127, 129]. Next, we provide an overview of dyslexia screening related to this thesis.

### 2.2.1 Dyslexia Screening

Due to the fact that dyslexia has nothing to do with intelligence, the difficulties for reading and writing seem to be just one thing they are not so good at. However, as mentioned in the previous section, dyslexia is connected to various factors instead of one cause [27].

Historically, the rates of spelling mistakes and reading errors have been the most common way to detect persons with dyslexia, using the widely-used paper and pencil assessments. Examples of the most common ways to detect children with dyslexia are summarized in [148]. Since people with dyslexia exhibit higher reading and spelling error rates than people without dyslexia independent of language (e.g., English [23], German [138], Greek [102]), there are diagnoses of dyslexia based on the error scores [140]. For instance, dyslexia is diagnosed in Germany if the spelling performance of the child is significantly under the level expected for a child of his or her age and general intelligence (example: 10 error words out of 40 words [138]). Other causes, such as insufficient vision or hearing ability, brain injury, or lack of opportunity to learn, must be excluded first [163, 164]. To confirm the diagnosis of dyslexia, there has to be a considerably high intelligence performance relative to a low spelling performance.

Most current approaches to detect dyslexia require linguistic skills (*i.e.*, phonological awareness or letter recognition for e.g., a spelling test [51]), expensive personnel (*i.e.*, psychologists) or special hardware (*i.e.*, MRI or fMRI Scans [94, 95, 150] or eye-tracking [4]). Various tools to detect dyslexia exist [148], but these detection tools are paper-based, using pen and paper assessment as well as indicators related to linguistic skills.

Detection and especially early detection of dyslexia is important because early intervention avoids adverse effects of dyslexia such as school failure and low self-esteem [138]. Often, children and their families have already experienced failures and frustration due to the inexplicable problems with learning how to read and write. Spelling tests force children with dyslexia to fail again while under observation, leading to additional stress and frustration. Therefore, in recent years, computer games have been used to provide support for children with dyslexia in an engaging, convenient, and cost-efficient way [75, 107, 128, 129].

There are few applications, especially web applications, for playful and low-cost screening; hence, it is an emerging field of research. We then present the state-of-the-art in the next sections on different approaches for screening readers and pre-readers using web applications and/or games.

## 2.2.2 Screening for Readers

Different spelling tests and reading tests are provided in different languages and have their history in paper-based detection. Paper-based spelling tests, such as the Diagnostische Rechtschreibtest (DRT) [51] in German, the PROLEC in Spanish [24, 25], and the DST-J in English [36], focus on the spelling and reading mistakes of children compared to their peers.

Tools and research focus on differentiating a person with dyslexia from a person without dyslexia (criteria of [2, 163, 164]). Various approaches to screening with a web application or analysis with machine learning have been successful for different use cases: machine learning prediction of eye-tracker fixation points when reading [4]; assessment of standardized tests battery in a web application using reading and attentional tests [11]; support vector machine (SVM) classification using content from electroen-

cephalography (EEG<sup>1</sup>) [39]; or the prediction with game measures using machine learning with content related to linguistic and attentional abilities [127]. A survey of machine learning approaches to distinguish readers with and without dyslexia concluded that dyslexia has to be differentiated by a language-based classification [6]. Screening with applications for readers is mainly based on the perception of linguistic skills [11, 79, 86, 127] (e.g., phonological awareness, letter recognition) but also on visual or auditory short-term memory [127] or phonological processing [133]. Mainly, these web applications have been designed as a low-cost approach for non-professionals as a quick screening tool to identify people that may have dyslexia and should go to see a professional. An overview of the cognitive skills tested, using tools related to this thesis, is given in Tables 2.1 to 2.3.

Next, we explain in detail those tools for detecting children with dyslexia.

- **Dytective** [125, 127, 129] is a web-based game with different stages to detect dyslexia with machine learning prediction models. The stages exist in German, English, and Spanish. Each stage has a new task, e.g., click the target non-word in a grid (see Figure 2.1, a) or search for the different letter in a letter grid (see Figure 2.1, right). *Dytective* in English has an accuracy of 83% for detecting a person with dyslexia ( $\sum = 267$ ;  $n = 52$  with dyslexia,  $n = 9$  maybe with dyslexia and  $n = 206$  without dyslexia) [129]. *Dytective* in Spanish has an accuracy of almost 85% in a small data set ( $\sum = 243$ ;  $n = 95$  with dyslexia,  $n = 31$  maybe with dyslexia and  $n = 117$  without dyslexia) [127]. In addition, a recent study with 4,333 participants was shown to have an accuracy of around 80% for detecting a person with dyslexia [125]. As of yet, we could not find a published evaluation for German. We report in the Table only the results from the Spanish evaluation with 4,333

---

<sup>1</sup>EEG monitors the brain activity.

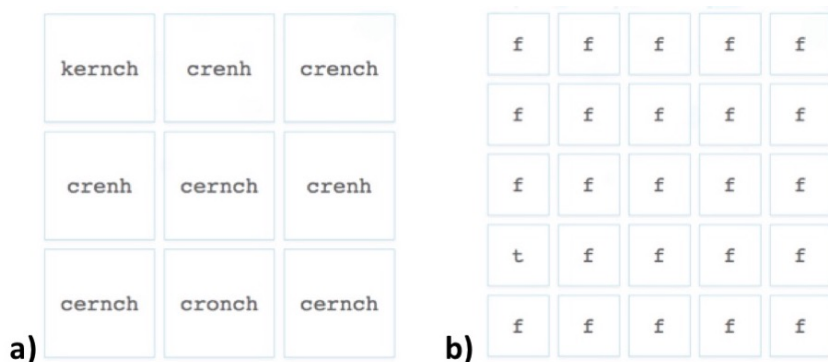


Figure 2.1: Example exercises from the dyslexia screener *Dytec-tive English*: **a)** player needs to click the target non-word listed among the distractors; **b)** player needs to click on the different letter [129].

participants. The most informative features at the individual level are how many correct and incorrect answers a participant has. They plan to include other languages in the future.

- **GraphoGame** [80] is a game to teach and evaluate early literacy skills. From their pre-analysis of children at risk in Finnish, they focus on delayed letter knowledge. Measurements are, for example, phonological manipulation, naming speed, or verbal short-term memory. GraphoGame provides exercises for children aged two to six. It requires reading skills, which is why we included it in this section on screening for readers, although kindergarten children were part of a longitudinal evaluation. We could not find a published prediction accuracy for screening children with dyslexia, but the product website states various use cases and research publications for the intervention, also in different languages [50].

<b>Tools</b>	<b>Dydetective</b> 2016, 2018 [125, 127, 129]	<b>GraphoGame</b> 2015 [80]	<b>Lexercise</b> 2016 [79]	<b>Nessy</b> 2014 [20]
<b>Languages</b>	Spanish English	Finnish	English	English
<b>Duration</b>	10 – 15 min.	n/a	n/a	~20 min.
<b>Skill</b>				
<i>Memory Skills</i>				
General	✓	✓		✓
Working memory	✓	✓		✓
Visual word memory	✓	✓		✓
Visual sequential memory	✓			✓
Visual alphabetical memory	✓	✓		
Auditory sequential memory	✓			✓
Auditory phonological memory	✓	✓		
<i>Processing speed</i>		✓		✓

Table 2.1: Cognitive skills tested in different dyslexia screening tools for readers (part one).



<b>Tools</b>	<b>Dyctective</b> 2016, 2018 [125, 127, 129]	<b>GraphoGame</b> 2015 [80]	<b>Lexercise</b> 2016 [79]	<b>Nessy</b> 2014 [20]
<i>Language skills</i>				
General	✓	✓	✓	
Alphabetic awareness	✓	✓		
Lexical awareness	✓	✓		
Morphological awareness	✓	✓		
Phonological awareness	✓	✓		✓
Semantic awareness	✓			
Syllabic awareness	✓	✓		
Syntactic awareness	✓		✓	
<i>Executive functions</i>				
General	✓			
Activation and attention	✓			
Sustained attention	✓			
Simultaneous attention	✓			

Table 2.2: Cognitive skills tested in different dyslexia screening tools for readers (part two).

<b>Tools</b>	<b>DyTECTive</b> 2016, 2018 [125, 127, 129]	<b>GraphoGame</b> 2015 [80]	<b>Lexercise</b> 2016 [79]	<b>Nessy</b> 2014 [20]
<b>Study</b>	Quantitative	Quantitative	No study	Quantitative
<i>Analysis</i>	Machine learning	Machine learning		Multiple regression
<i>Participants total</i>	4,333	267		69
<i>With dyslexia</i>	469	52		
<i>Maybe with dyslexia</i>		9		
<i>Without dyslexia</i>	3,864	206		
<i>Age</i>	7 – 70	7 – 60		7 – 15
<i>Accuracy</i>	80%	85%		

Table 2.3: Evaluation of reader screening tools.

- **Lexercise Screener** [79] is an English screening tool for detecting dyslexia. Children read familiar words, and the parent records the child's response. This requires phonological awareness from the parents, which might be difficult if the parents have been diagnosed with dyslexia themselves. Hence, a lack of objectivity needs to be taken into account.
- **Nessy** [20, 86] is another English screening tool for detecting dyslexia. Exercises are designed to test many cognitive skills (see Table 2.1). It provides exercises for children aged five to sixteen years, and the test takes around 20 minutes. The research summary published on their website reports the results of the multiple regression analysis between the game and the *comprehensive test of phonological processing* [18], with a strong correlation of almost 0.8.

All of the web applications above are mainly designed for desktop or laptop computers or tablets. We could only find a published prediction accuracy for *Dytective*. *Nessy* has been evaluated with multiple regression analysis, but we could not find the details of the study apart from the *research brief* published on its website. The reason might be that these applications are products, and therefore their approaches are trade secrets.

To sum up, all these screening applications are language dependent. This means, on one hand, that the content of the application needs to be adapted for every new language, which is time and resource consuming. On the other hand, only people who have already acquired language can be tested (*i.e.*, children need a minimum knowledge of phonological awareness, grammar, and vocabulary for the application to detect or predict dyslexia). In practice, these tools can only screen children after the first year of school and not earlier. Therefore, new ways of detecting the risk of having dyslexia are needed for pre-readers. Next, we present the results of our literature review for pre-reader screening approaches.

### 2.2.3 Screening for Pre-Readers

The detection of dyslexia in children before they learn to read and write is difficult because the obvious indicators manifest themselves mainly in reading and writing. As we already pointed out in the previous sections, the difficulty in detecting dyslexia before children go to school is in the missing phonological awareness and linguistic abilities. This means that children can be detected only after they begin to learn how to read and write. This puts students with dyslexia behind. Therefore, new ways of detecting the risk of having dyslexia are needed for pre-readers.

Prior studies show expensive approaches to predict future language acquisition of pre-readers, *e.g.*, from newborns with brain recordings [52, 80], infants with *rapid auditory cues* [8] to kindergarten children with the perception of *visual-spatial attention* [38]. Previous research has related speech perception difficulties to auditory processing, phonological awareness, and literacy skills [133, 149]. Phonological deficits of dyslexia have been linked to basic auditory processing [53]. The auditory perception of children with dyslexia has been proven to be related to the sound structure [63] as well as to the auditory working memory [82]. Neither of these requires reading ability, and thus they may be useful in detecting dyslexia.

Related research suggests that reading difficulties are due to problems with visual-spatial attention and poor coding instead of phonological difficulties [156]. Apart from that, visual discrimination and search efficiency are used as predictors for future reading acquisitions [38].

Now, we present selected examples to explain different approaches that aim to predict dyslexia in pre-readers. We focus on digital, playful, and low-cost pre-reader screening. All of them base their approaches on indicators mainly related to linguistic skills. This rationale is supported by the following assumptions: (1) dyslexia does not develop when children come to school, but is

already there before; (2) linguistic related indicators can represent the difficulties a person with dyslexia has with writing and reading; and (3) dyslexia can be measured through the interaction behavior of a person. An overview of the cognitive skills tested in each application is given in Table 2.4 and their evaluation are presented in Table 2.5 if results are published. The applications are:

- **AppRISE** [151] is a recently published promising concept to screen pre-readers for dyslexia, although the design and development as well as the validation are not yet published. This is the reason we only consider the descriptions from the lab research website [151] and the game website [100, 101]. The goals are to develop a screening tool with non-linguistic cues and tutorials for English learners [100]. Besides the cognitive skills presented in Table 2.4, the tablet-based games aim to measure strength and weakness using literacy modules (e.g., *receptive vocabulary*) and cognitive modules (e.g., *cognitive flexibility or nonverbal reasoning*) [100]. The full assessment takes around 3 hours and can be split into sessions of 30 minutes on different days and a validation study should be currently in progress [101].
- **AGTB 5–12** is a computer-based test for children from age five to twelve years old [55, 56, 66]. In Germany, it was one of the first applications that addressed the visual and phonological working memory (quoted after [66]), in addition to the linguistic skills and working memory. An evaluation with over 1,659 children was conducted in 2008 and 2009 (quoted after [66]). Based on the results of the pilot tests, which also included pre-readers (age 4 or 5), the final AGTB aims to screen children from ages 5 to 12 [56]. We could not find the details of this study. On the product website, it is stated that the Cronbach Alpha is between .58 and .98 for children age five to eight, while is .67 and .99 for children from the age of nine and older [56]. *AGTB 5–12* is criticized for its lack of objectivity for

<b>Tools</b>	<b>AppRISE</b> 2019 [151] English 3 hours	<b>AGTB 5–12</b> 2012 [56, 66] German ~ 87 min.	<b>BELS</b> 2018 [40, 41, 42] n/a 20-30 min.	<b>DYSL-X</b> 2013 [45, 152] n/a n/a	<b>GC</b> 2017 [44] Italian endless	<b>Lexa</b> 2018 [98] English n/a
<b>Skills</b>						
<i>Memory general</i>	✓	✓				✓
Short-term memory		✓				
Auditory phonological memory		✓	✓			✓
<i>Processing speed</i>	✓					✓
<i>Language skills</i>						
general	✓	✓	✓	✓		
Alphabetic awareness	✓			✓	✓	
Phonological awareness			✓		✓	

Table 2.4: Cognitive skills tested in dyslexia screeners for pre-readers.

<b>Tools</b>	<b>AGTB 5 –12</b>	<b>DYSL-X</b>	<b>GC</b>	<b>Lexa</b>
<b>Study Analysis</b>	2012 [55, 56]	2013 [45, 152, 153]	2017 [43, 44]	2018 [98]
<b>Participants total</b>	Quantitative Psychological testing 1,659	Qualitative, Quantitative Empirical evaluation, UX Testing e.g., 15 and 20	Qualitative, Quantitative Usability Test 23	Quantitative Classification 56
<b>With dyslexia</b>			6	20
<b>Maybe with dyslexia</b>			17	36
<b>Without dyslexia</b>				
<b>Age</b>	5 – 12	Pre-reader	3 – 6	6 – 15
<b>Results</b>	Cronbach Alpha .58 and .98 (age 5 – 8) and .67 and .99 (age > 9)	Pre-readers prefer intuitiveness and physicality of their hands, e.g., touch input	Found differences between children at risk and the control group, e.g., game score, number of won matches	Accuracy of 89.2% (all features) and 53.8% only phonological processing

Table 2.5: Study details pre-reader screening tools if results are published.

some tasks because the supervisor has to decide the grading depending on subjective knowledge [66]. Although *AGTB 5–12* aims to screen pre-readers, the graphical user interface and its interaction is not specifically designed for smaller children. Additionally, both the long duration of over an hour and the detailed instructions are not suitable for younger children.

- **BELS** (Boston Early Literacy Screening) aims to screen dyslexia from four years old. The screener is a recently published promising concept, though the design and development as well as the validation are not yet published. This is why we only consider the descriptions from the lab research website [40, 42] and the game website [41]. In addition to the cognitive skills presented in Table 2.4 is the gamified screener using language related indicators: *rapid automatized naming, letter (sound) knowledge, vocabulary and oral listening comprehension* [42]. The validation study should be currently in progress.
- **DYSL-X** (also called DIESEL-X) aims to predict the possibility of a child having dyslexia at the age of five [45, 152]. The three mini-games are designed to measure dyslexia using, for example, indicators such as letter knowledge, frequency modulation detection, and end-phoneme recognition [45]. Participants play each of the three games (for an example, see Figure 2.2, *d*) four times with no time limit. An example task is finding a letter. Various empirical evaluations on different prototypes were conducted to improve the gameplay and future implementations for pre-readers, *e.g.*, 15 pre-readers participated in a diary study and 20 pre-readers participated in interviews [45]. The results show that pre-readers prefer *touch input over keyboard or mouse* [45].
- **Game–Collection** (CG) has six games, each with a different challenge and gameplay [43, 44]. The games use visual and auditory elements and an evaluation was done only on



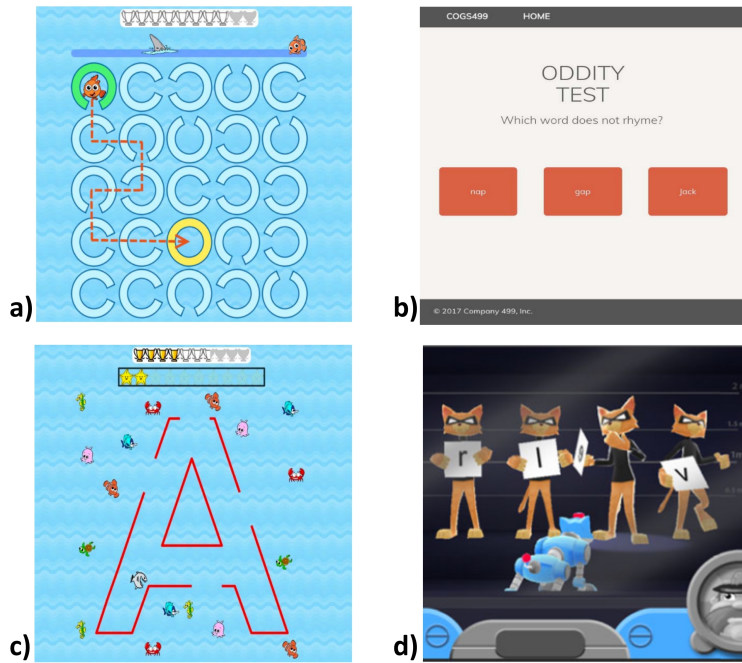


Figure 2.2: Screen examples: **a)** *Paths* game [43]; **b)** *Lexa* [98]; **c)** *Fence letters game* [43]; and **d)** *DYSL-X* [152].

the game interaction. The games explore visual cues and temporal perception for predicting dyslexia at the age of five or six [43, 44], although children of age three or four tested the games as well. In the game called *Paths*, a shape with similarities to the letter *C* is used as an indicator (see Figure 2.2, *a*). The game called *Fence Letters*, tries to distract a child while they close the lines to create a letter (see Figure 2.2, *c*). The usability test reported that children without dyslexia ( $n = 17$ ) got a higher game score, winning a higher number of matches and using less time than children with risk of dyslexia ( $n = 6$ ).

- **Lexa** [98] is a prototype to detect dyslexia via auditory processing using oddity and rise time (see Figure 2.2, *b*). The simple decision tree analysis of the lab study data (data was collected by Goswami *et al.* [46]) was used to find the most relevant features. With the MATLAB classification, a higher accuracy (89.2% vs. 53.8%) was found if no pre-processing of the feature related to phonological processing was applied. The features description is missing, although various tests are mentioned, such as the Wechsler Intelligence Scale for Children (WISC) or the British Ability Scales (BAS), to collect features. The small sample size ( $n = 56$ ) with a  $\sim 90\%$  accuracy could mean that the model is over-fitting and the classification model is only accurate for this data set because no further measures such as confusion matrix, recall, F1-score or precision are mentioned, and no validation is performed. The researchers state that the biggest challenge is creating different rise times sounds and determining whether a child is guessing the answer.

The web games presented focus mainly on high-score game play, easy instructions, colorful representation, and story based design. *AGTB 5–12* [55, 148] and *Lexa* [98] predict the risk of dyslexia. Besides *AGTB 5–12* [55, 148], all games are prototypes and have not been brought to the market until now. Only *Lexa* [98] reports an accuracy (89.2%) using features related and not-relevant to phonological processing. However, the tool is not playful, and features are collected with extensive tests which probably take considerable time and cost. In addition, the classification of machine learning is carried out on a small sample ( $n = 56$ ), without any validation and with no precautions or discussions about over-fitting. Apart from *Lexa* [98], so far, no evaluation for the prediction accuracy of any of these games has been made public. Also, the focus for the prediction is mainly on letter knowledge and phonological awareness.

The GaabLab collected an exhaustive list of dyslexia screeners [93] mainly for pre-readers, and there are various new applications and approaches. Examples of games and approaches are: a clinical machine learning analysis for language disorders ages 3 to 5 [72]; a serious game using gesture-based interfaces for clinical diagnosis of learning difficulties [22]; or a concept and puzzle game evaluated with a usability test [111].

## 2.2.4 Auditory Perception

As mentioned in the previous sections, dyslexia has been connected to various auditory indicators and empirical results. We provide in this section an overview of studies and indicators related to auditory perception, which are essential for our thesis.

From the analysis of error words, we know that children have difficulties with ambiguous words [118, 126] when the language has phonetic ambiguity. For example, Greek has almost no phonetic ambiguity because a grapheme refers to a phoneme, which results in fewer phonological errors. French, on the other hand, with its higher phonetic ambiguity, has more phonetic errors [102].

The central assumption of the hypothesis *dyslexia is connected to auditory perception* is that phonological indicators or deficits represent dyslexia [27, 46]. Various studies support the hypothesis because they found evidence in dyslexia deficits of phonology such as auditory processing [53], prosodic skills [49], phonemic awareness [49] or phonological awareness (quoted after [27]).

For example, the *rapid auditory processing deficit hypothesis* assumes that individuals with dyslexia have problems processing short auditory cues. Another theory claims the dynamic change of the acoustical parameters causes the difficulties [53].

The *rise time theory* suggests a connection between dyslexia and slow auditory procession or impaired discrimination of amplitude [27]. Since the *phonological grammar* of music [99] is similar to the prosodic structure of language, music, *i.e.*, a combination

of acoustical parameters, can be used to imitate these features. Studies showed a significant difference in the perceptions of readers with dyslexia on the syllable stress compared to those of the control group at the age of 9 [49].

For example, the rise time of a sound could imitate stress levels on syllables. Additionally, findings suggest a relation between rise time perception and *prosodic* and *phonological development* [63]. Even newborns respond automatically to the complex task of perceiving music [165] and show differences in perception of *sensitivity to native versus non-native rhythmic stress* [47] by the age of 5 months or to the phonemic length by 6 months [80]. Because of the similarities of music and language, different acoustic parameters of sound have been explored and proven significant in the perception of children at the age of 8 to 13 years with and without dyslexia [63] e.g., *rise time*, *short duration (100ms)*, *intensity*, and *rhythm*. Also, the perception of pitch and its patterns relates to reading skills, which is a main area of difficulty for people with dyslexia [133, 165]. Furthermore, a lab study showed different behavior of infants ( $m = 7.5$  months) using complex sound frequencies to predict a child's language skills at the age of three [8].

A recent study found evidence that dyslexia-associated genes are related to the encoding of sounds in the auditory brainstem [9]. However, there are musicians with dyslexia who scored better on auditory perception tests than the general population [82]. At the same time, these participants score worse on tests of auditory working memory, *i.e.*, the ability to keep a sound in mind for seconds. This observation is in line with the results on perceptions for short duration sounds [63] and the findings on the *prosodic similarity effects* of participants with dyslexia [47]. One connection between the difficulties in perception of language and music seems to be the problem with short-term memory and recall of information chunks [47]. Since people with dyslexia have short-term memory difficulties [71, 92], questions like "*Which sound did you hear first*" or "*Which sound is pitched higher?*" [63] could deter-

mine the groups. Huss *et al.* [63] already showed that significant performance differences can be found between children age 8 to 13 with and without dyslexia using musical metrical structure in a controlled setting. These auditory indicators, such as phonological differences, could be used in a language-independent approach to screen for dyslexia.

## 2.2.5 Visual Perception

Apart from the auditory perception results already mentioned in the previous section, previous research suggests that the cause of reading difficulties could be partly due to lack of *visuo-spatial* (also called *visual-spatial* [38]) attention and poor visual coding instead of auditory difficulties [156]. This would mean that the difficulties people with dyslexia have in reading and writing are due to a poor decoding of visual cues, e.g., letter recognition, especially for error cases where a person has a good phonological awareness but difficulties in reading non-words.

Further, findings also provide evidence that the cause of dyslexia might be due to a *more basic cross-modal letter-to-speech sound integration deficit* and *pre-reading visual parietal-attention* [38]. These can predict reading acquisition in preschoolers with visual-spatial attention. An example of a visual-spatial attention task is a search task (searching for symbols), which shows significant differences in the error rate for poor readers in first grade compared to their peers [38].

The analysis of error words from children with dyslexia shows that the correct and incorrect letters in error words are visually similar, which holds true for different languages, e.g., English, Spanish [126] or German [118]. The annotated error and correct letters show similarities in different visual features called *mirror letter* (e.g.,  $\langle n \rangle$   $\langle u \rangle$ ) or *fuzzy letter* (e.g.,  $\langle s \rangle$  and  $\langle z \rangle$ ). Some letters have also similarities in the *vertical* (e.g.,  $\langle m \rangle$ ) and *horizontal* symmetries (e.g.,  $\langle E \rangle$ ) through their visual features [126].

These visual indicators, such as horizontal or vertical symmetry of visual representation, could be used in a language-independent approach to screen for dyslexia.

## 2.3 Design

An interdisciplinary thesis such as ours requires a standardized approach in order to allow other researchers to evaluate and interpret our results. Therefore, we used the *design science (DS) research methodology (DSRM)* [96] and the *human-centred design (HCD)* [67], which are both well-known and well-defined design concepts. We describe the core elements for each design process in the next sections. Additionally, we explain the main concepts behind our experimental design and data analysis.

### 2.3.1 Design Science Research Methodology

Researchers combine methodologies or approaches and need to evaluate results from other disciplines. Combining discipline techniques is a challenge because of different terms, approaches, or communication within each discipline. For example, the same term, such as *experiments*, can have a different interpretation in *data science* versus *human computer interaction (HCI)* approaches. In HCI, *experiments* mainly refer to user studies with humans, whereas in data science, experiments refer to running algorithms on data sets.

The *design science research methodology (DSRM)* supports standardization of design science, for example, to design systems for humans. The DSRM provides a flexible and adaptable framework to make research understandable within and between disciplines [96]. Since the early 1990s, design science is integrated into information systems and provides with the DSRM a methodology to justify system design research (quoted after [96]). The core

elements of DSRM have their origins in human-centred computing and are complementary to the human-centred design framework [61, 67]. DSRM suggests the following six steps to carry out research: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication.

In the first step, researchers describe the problem and the motivation for the technological solution. The level of detail depends on the complexity of the problem. Next, the goals and functionality of the solution are stated, taking into account the information from the previous step and quantifying the solution. In step three, a technological solution is designed and implemented with the proposed architecture or functionality. In the next two steps, the technological solution is presented and evaluated to compare with the goals set at step two. In the last step, researchers “*communicate the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences, such as practicing professionals, when appropriate*” [96].

The information system design theory can be considered to be similar to social science or theory building [158]. However, designing systems was not and is still not always regarded to be as valuable research as “*solid-state physics or stochastic processes*” [147]. One of the essential attributes for design science is a system that targets a new problem or an unsolved or otherwise important topic for research (quoted after [96] and [58]). If research is structured in the six steps of DSRM, a reviewer can easily analyze it by evaluating its contribution and quality. In addition, authors do not have to justify a research paradigm for system design in each new thesis or article.

### 2.3.2 Human-Centred Design

The *human-centred design* (HCD) framework [67] is a well-known methodology to design interactive systems that takes the whole de-

sign process into account and can be used in various areas: enterprise software [105, 110], health related applications [1, 43, 108, 152], remote applications (Internet of things) [135], social awareness [157], or mobile applications [1, 111]). With the HCD, designers focus on the user when developing an interactive system to improve *usability* and *user experience*.

*“Human-centred design is an approach to **interactive systems development** that aims to make systems **usable and useful by focusing on the users** [...] and by applying human factors/ergonomics, and usability knowledge and techniques. This approach enhances effectiveness and efficiency, **improves human well-being**, user satisfaction, **accessibility and sustainability**; and counteracts possible adverse effects of use on human health, safety and performance.”* ISO 9241-210 [67]

The two main terms to describe and quantify the methods for HCD are *usability* and *user experience* (UX). How a user interacts for a certain goal or task in a specific context is called *usability* [68]. This means a certain type of user (for example a student) wants to do a specific task (for example, writing an email to her/his professor on her/his computer from home). The level of detail for the task description can depend on the design resources, (*i.e.*, time or personnel) or design goal, (*i.e.*, proof of concept or product). The main focus is on the user achieving the task effectively, efficiently, and with satisfaction. *User experience* incorporates usability and advances the concept of interaction through the perception and responses of the user as well as the “*emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors and accomplishments that occur before, during and after use*” [67].

The HCD is an iterative design process (see Figure 2.3). The process starts with the planning of the HCD approach itself. After that, the (often interdisciplinary) design team members (*e.g.*, UX designers, programmers, visual designers, project managers



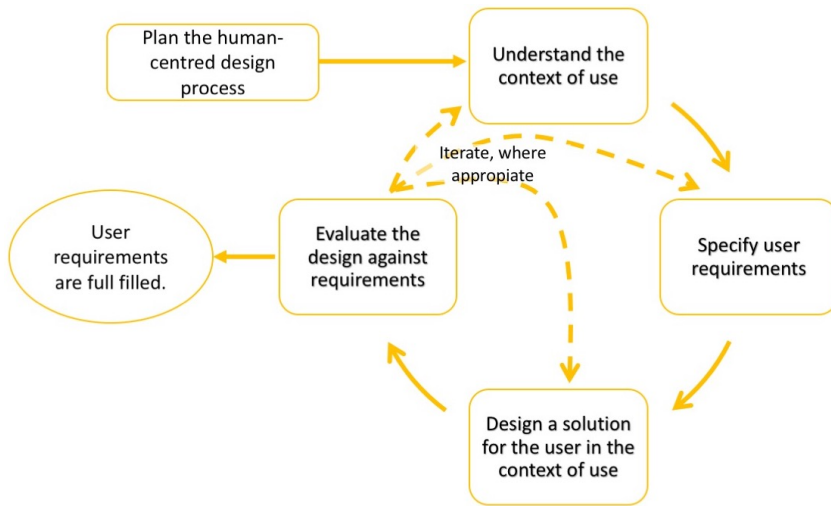


Figure 2.3: Activities of the human-centred design process visualised by the author from [68].

and/or scrum masters) define and understand the context of use, (e.g., at work in an open office space). Next, user requirements are specified and can result in a description of the user requirements or a *persona* to communicate the typical user's needs to e.g., the design team [7]. Subsequently, the system or technological solution is designed with the defined scope from the context of use and user requirements. Depending on the skills or the iterative approach, the designing phase can produce a (high- or low-fidelity) prototype or product as an artifact [7]. A low-fidelity prototype, such as a paper prototype, or a high-fidelity prototype, such as an interactive designed interface, can be used for an iterative evaluation of the design results with users [3].

Ideally, the process finishes when the evaluation results reach the expectations of the user requirements. Otherwise, depending on the goal of the design approach and the evaluation results, a new iteration starts either at understanding the context of use, specifying the user requirements, or re-designing the solution. Early and

iterative testing with the user in the context of use is a core element of the HCD. This is especially true for new and innovative products, as both the scope of the context of use and the user requirements are not yet clear and must be explored.

There are various methods and artifacts which can be included in the design approach depending, for example, on the resources, goals, context of use, or users. An example for a design method is in the principles of *Gestalt psychology* [91]. The principles are especially useful when the user interaction cannot be known from previous experience. The design principles of *affordance* or *law of proximity* [91] support the design of new technological solutions. Affordance describes how the interaction is known by the design, e.g., how to use a door handle when it is designed differently. Law of proximity describes how close design elements belong together, e.g., button and description.

Evaluation methods are, for example, the five-user study [89], the User Experience Questionnaire (UEQ) [59, 112, 119], observations, interviews, or the think-aloud protocol [16]. For instance, the five-user study is commonly used to find the major difficulties in the prototype or product, which is iterated over the designed solutions [89, 91]. This is especially helpful for new technological solutions with limited resources. The five-user test can also be divided by user groups, e.g., children, parents, students, power users, or routine users.

Methods can be combined to get quantitative and/or qualitative feedback, and the most common sample size at the Computer Human Interaction Conference (CHI) in 2014 was 12 participants [19]. With small testing groups ( $n < 10 - 15$ ) [19], mainly qualitative feedback is obtained with (semi-structured) interviews, think-aloud protocol, or observations. Taking into account the guidelines for conducting questionnaires by rules of thumb, like the UEQ could be applied from 30 participants to obtain quantitative results [120].

### 2.3.3 General Considerations of Research Design

In a quasi-experimental study, there is control over certain variables such as participant attributes, which then assigns participants to either the control or the experimental group [37]. An example of such an attribute could be whether or not one has a dyslexia diagnosis. In a *within-subject design* [37], all participants take part in all study conditions, e.g., tasks or game rounds. When applying a *within-subject design*, the conditions need to be randomized to avoid *systematic or order effects* produced by order of the conditions [37]. These unwanted effects can be avoided by counterbalancing the order of the conditions, for example with Latin Squares [37].

The advantage of a *repeated-measures design* in a *within-subject design* is that participants can engage in multiple conditions [37]. When participant attributes such as age or gender are similar in different groups, a repeated-measures design is more likely to reveal the effects caused by the dependent variable of the experiment. When conducting a *within-subject design* with a repeated measures design, and assuming a non-normal distribution for independent participant groups, application of the *Mann-Whitney-Wilcoxon Test* (also called *independent Wilcoxon Test*) is recommended [37]. To avoid confusion with the *Wilcoxon Signed-Rank Test*, which is for dependent groups, we will refer to the *Mann-Whitney-Wilcoxon Test* as simply the *Mann-Whitney U Test*, as it has been called by various authors. [37, 146].

Testing for one hypothesis, such as “a *person with dyslexia clicks more than a person without dyslexia*”, involves a *one-tailed test* [37]. The *homogeneity of variance* assumes that the variation of a population in different experimental conditions is nearly equal.

The *American Psychological Association* recommends integrating the *effect size* to estimate the size of effect for the population (quoted after [37]). To quantify the effect of the results, the *effect size* can be calculated with, for example, Cohen’s *d*. Or, the effect size (*r*) can be calculated as

$$r = \frac{z}{\sqrt{N}}$$

where  $z$  is the  $z$ -score and  $N$  is the number of observations [37]. The effect size is considered to be small if the value is 0.10, medium if it is 0.30, and large if it is 0.5 [37]. As for psychology in HCD, multi-variable testing must be addressed to avoid having significance by chance. This can be achieved by using a method such as *Bonferroni-Correction* and having a clear hypothesis.

*Big data* has a different meaning to people depending on the research context, profession or mindset. We use the term “*big data*” in terms of size [34]. The decision for an algorithm to investigate your data set depends on the size, quality and nature of the data set as well as the available computational time, the urgency of the task, or the research question. In some cases, small data is preferable to big data because it can simplify the analysis [34]. In some circumstances, this leads to more reliable data, lower costs, and faster results. In other cases, only small data is available.

Dependent measures are used to find, for example, differences between variables, [37] while features are used as input for the classifiers to recognize patterns [13]. Machine learning is a data-driven approach in which the data is explored with different algorithms to minimize the objective function [30]. We are referring to the implementation of the Scikit-learn library version 0.21.2 if not stated otherwise [143]. Although a hypothesis is followed, optimizing the model parameters is not generally considered problematic unless we are *over-fitting* (also written as *overfitting*), as stated by Dietterich in 1995:

*“Indeed, if we work too hard to find the very best fit to the training data, there is a risk that we will fit the noise in the data by memorizing various peculiarities of the training data rather than finding a general predictive rule. This phenomenon is usually called overfitting.”* [30]

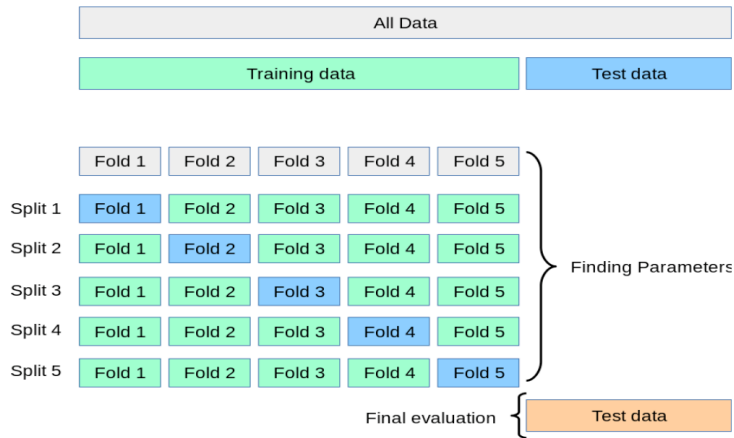


Figure 2.4: Approach of cross-validation from [141].

If enough data is available, common practice *holds out* (that is separating data for training, test or validation) a percentage to evaluate the model and to avoid over-fitting, e.g., a test data set of 40% of the data [141]. A validation set (holding out another percentage of the data) can be used to, say, evaluate different input parameters of the classifiers to optimize results [141], e.g., accuracy or F1-score. Holding out part of the data is only possible if a sufficient amount of data is available. Models performed on small data are prone to develop over-fitting due to the small sample and feature selection [69]. Cross-validation with k-folds can be used to avoid over-fitting when optimizing the classifier parameters (see Figure 2.4). In such cases, the data is split into training and test data sets. A model is trained using subsets (typically k-folds, 5-folds, or 10-folds) of the training set, and is evaluated using the test data [141]. It is recommended that one hold out a test data set while using cross-validation when optimizing input parameters of the classifiers [141]. However, small data sets with high variances are not discussed.

Model-evaluation implementations for cross-validation from Scikit-learn, such as *cross\_val\_score* function, use scoring parameters for the quantification of the quality of the predictions [142]. For example, with the parameter *balanced\_accuracy* imbalanced data sets are evaluated. The parameter *precision* describes the classifiers ability “not to label as positive a sample that is negative” [142]. Whereas the parameter *recall* “is the ability of the classifier to find all the positive samples” [142]. As it is unlikely to have a high precision and high recall, the *F1-score* (also called F-measure) is a “weighted harmonic mean of the precision and recall” [142]. Scikit-learn library suggests different implementations for computing the metrics (e.g., recall, F1-score) and the confusion matrix [141]. The reason is that the *metric function* reports over all (cross-validation) fold, whereas the *confusion matrix function* returns the probabilities from different models.

## 2.4 Gamification

*Gamification* has been successfully used in various use cases and applications and frameworks have been established [54, 84, 132, 144]. The concept of *Gamification* is using game elements related to, for example, emotions or progressions in applications for different contexts to engage and motivate users. Gamification can be used to design the *gameplay* of a game. The *gameplay* can be described as “the degree and nature of the interactivity that the games include” [134]. This can refer to, for example, the rules within which players can make their choices, such as deciding to interact with another game character. The gameplay is what allows the interaction of the player with another game character. However, the player makes the choice of agreeing or disagreeing in the conversation, which determines how the game reacts. To support the design of educational environments that improve students’ engagement and motivation, applying the concept of gamification is beneficial.

## 2.4.1 The Evolution of Gamification

At first, in 2003, the term gamification was used to describe the process of making electronic devices more entertaining [160]. Gamification appeared in the domain of software engineering in 2008 and gained widespread acceptance in the following years, significantly influenced by the presentation of Schell [137].

The marketing-based perspective from Zichermann and Cunningham [166] defines gamification as a process of applying game-thinking and game mechanics to engage users and solve problems. Game mechanics are defined as various game elements such as points, levels, leaderboards, badges and challenges.

Since 2011, the concept of gamification has received comprehensive academic attention as different approaches to defining the concept have emerged. Deterding et al. [28] define gamification as the use of game design elements in non-game contexts. They define game design elements as components, patterns, guidelines, and models, as well as processes and practices from the field of game design. The design aspect of the definition is used to delimit the creation of gamified applications in relation to fully fledged games, especially for entertainment purposes. The notion of a non-game context is proposed to define the application of gamification in contexts other than games. This definition makes no assumption about the purpose or field of application of gamification.

Huotari and Hamari [62] define gamification from a service marketing perspective as enhancing a service with affordances for gameful experience. They point out that the user's experience and value creation are influenced and supported by the gameful experience of a service. A service is described as any intentional act that helps an entity. The value of a service is determined by the user's individual perception, as the service provider can only propose an intended value.

The definition of Huotari and Hamari [62] stands in some contrast to the definition of Deterding [28]. Notably, it negates the

importance of explicit game elements for gamification. Rather, it states that gamification occurs by enhancing a service with any qualities that help to elicit a gameful experience.

Werbach and Hunter [160] portray a business oriented approach to gamification and define the concept as the use of game elements and game-design techniques in non-game contexts. The authors give a further definition of game elements, presenting a hierarchy of game elements in the form of a pyramid. This hierarchy separates game elements into levels of abstraction, with the most abstract (*dynamics*) at the top, *mechanics* in the middle, and concrete *components* at the bottom. They also introduce a six-step process for applying game elements to the target system which involves defining business objectives, identifying players and activities, and other actions.

Kapp [131] defines gamification as using game-based mechanics, aesthetics and game thinking to engage people, motivate action, promote learning and solve problems. At the core of the definition is the aspect of game thinking, which applies elements like competition, cooperation, discovery and narration to everyday activities.

From the contributions above, the definition of Deterding [28] is the most abstract and comprehensive, and thus it can be applied in many different contexts. Therefore, we use it to build a working definition upon which this work is based: gamification is the use of game-design elements in non-game contexts to enhance user experience.

## 2.4.2 Gamification vs. Serious Games

Gamification is difficult to define precisely, as shown by the many definitions above. As a tool in the field of software engineering, it touches on many different adjacent fields such as human-centred design or motivational psychology. It also stands in contrast to related concepts like the widely-known genre of serious games.



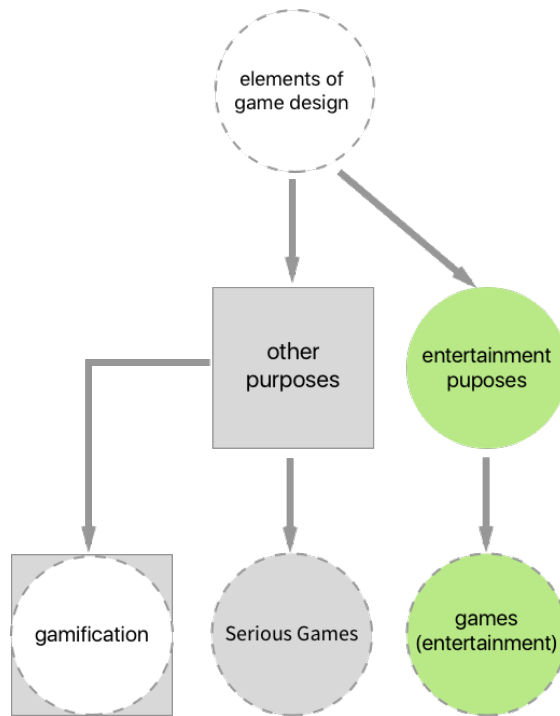


Figure 2.5: Our synthesis from elements of game design to different outcomes.

To explain the controversy, we shortly describe serious games. Sawyer [136] defines serious games as the meaningful use of computerized games whose primary purpose is not entertainment.

Serious games as a special form of video games have a long tradition, as there are examples as old as the first video games. Some of the most popular examples were created to serve a variety of purposes: The Oregon Trail, 1973, Education; America’s Army, 2002, Military; X-Plane 10, 2012, Training [76].

The idea of serious games stands in contrast to gamification, as gamification is exclusively applied to non-game systems. This distinction of terms is illustrated in Figure 2.5. The graphic shows the path a system can follow in its development. Start-

ing with the use of game design elements, the character of a system depends on the purpose the system pursues. The final status is determined by the user's perception of the system.

## 2.5 Summary

The previous sections presented the current knowledge of how to detect a child with dyslexia, auditory as well as visual perception, design, and games.

Dyslexia is a *specific learning disorder* [2, 163, 164], which is probably caused by the *phonological skills deficiencies associated with phonological coding deficits* [155] and problems with visual-spatial attention [38, 156].

In summary, there have already been improvements in both the screening of dyslexia and the evaluation of these approaches. However, detecting dyslexia in a pre-reader is especially challenging because dyslexia often causes reading and writing problems but does not show obvious deficits in other areas.

The tool *Lexa* [98] diagnoses dyslexia with an accuracy of 89.2% using features that are relevant and irrelevant to phonological processing. However, it uses various tests (extensive resources) to collect data and a small sample size analysis without a discussion of over-fitting.

Language-dependent content, extensive testing, and lack of precautions for over-fitting for small data motivated us to design a playful, language-independent game and collect a reasonable amount of data, while taking care of limited resources and taking precautions for over-fitting.

We aim to use auditory and visual cues as language-independent input within a game to collect dependent measures having less expense (*e.g.*, less tests, less personnel, less participants). The collected measures will be used as features for ma-

chine learning classifiers to achieve our main goal: accurate first prediction results for universal screening for the risk of dyslexia with language-independent content using games and machine learning.

We will perform user studies in collaboration with schools and associations for dyslexia from Germany and/or Spain. The *dependent variables* (also called dependent measures) that we plan to use will be derived from the interaction with online games, such as several performance measures (scores, misses, clicks, etc.).

In the next chapter, we describe our research questions and our approach to design a language-independent game that could also be used to provide pre-readers with more time to practice.

# Methodology

---

## 3.1 Introduction

An interdisciplinary thesis like ours requires a standardized approach to allow other researchers to evaluate and interpret our results. Therefore, we combine the *Design Science (DS) Research Methodology (DSRM)* [96] with the *Human-Centred Design (HCD)* [67]. In the following sections, we explain the integration of the DSRM and HCD as well as the other methodologies used in this thesis, which come from human-computer interaction, data science, and gamification. The contents of Section 3.4.3 were published in [122].

## 3.2 Combining HCD and DSRM

The *Design Science Research Methodology (DSRM)* provides six steps for carrying out research, which we answered with the *Actions* we plan to do and matched with the HCD phases (see Figure 3.1, black boxes). The blue boxes are only related to the DSRM, while the green boxes are also related to the HCD approach.

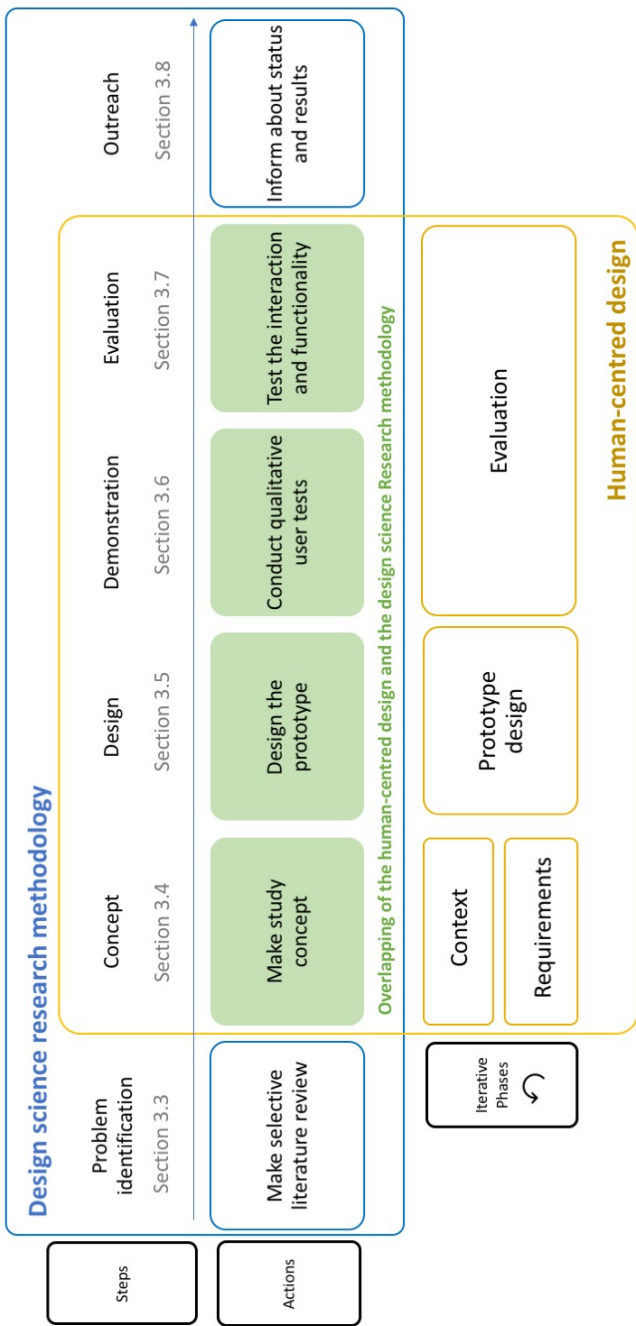


Figure 3.1: Integration of the *Human-centred Design* in the *Design Science Research Methodology*.

The four green boxes match the four HCD phases (see Figure 3.1, yellow boxes). Figure 3.1 shows how we approach each of the DSRM steps. First, we make a selective literature review to identify the problem, which is described in Section 3.3 and is further elaborated on in Chapter 2. This results in a concept for targeting the language-independent screening of dyslexia using games and machine learning, explained in Section 3.4. We then describe how we design and implement the content and prototypes in Section 3.5. How we test the interaction and functionality to evaluate our content (with the prototypes) is described in Section 3.6. Finally, Section 3.8 describes the outreach of our research.

We design our interactive prototypes to conduct online experiments with participants with dyslexia using the *human-centred design* [67]. The *human-centred design* complements the *design science research methodology* with a focus on the participants and also provides various guidelines, methods, and artifacts for the design of a prototype. For example, we use the following:

- an iterative design approach [68],
- a user requirements phase [68],
- evaluation methods such as (semi-structured) interviews or the *think aloud protocol* [16],
- principles from *gestalt psychology* [91], and
- artifacts such as high- and low-fidelity prototypes [16].

We present an overview of the HCD phases, our focus and our activities for each phase in Figure 3.2. With the HCD, we focus on the participant and the participant's supervisor (e.g., parent/legal guardian/teacher/therapist) as well as on the context of use when developing the prototype for the online experiments to measure differences between children with and without dyslexia. Our prototypes target the auditory and visual perception of children, which have been examined in lab studies already (selective literature review in Chapter 2) and which we adapt for online experiments.

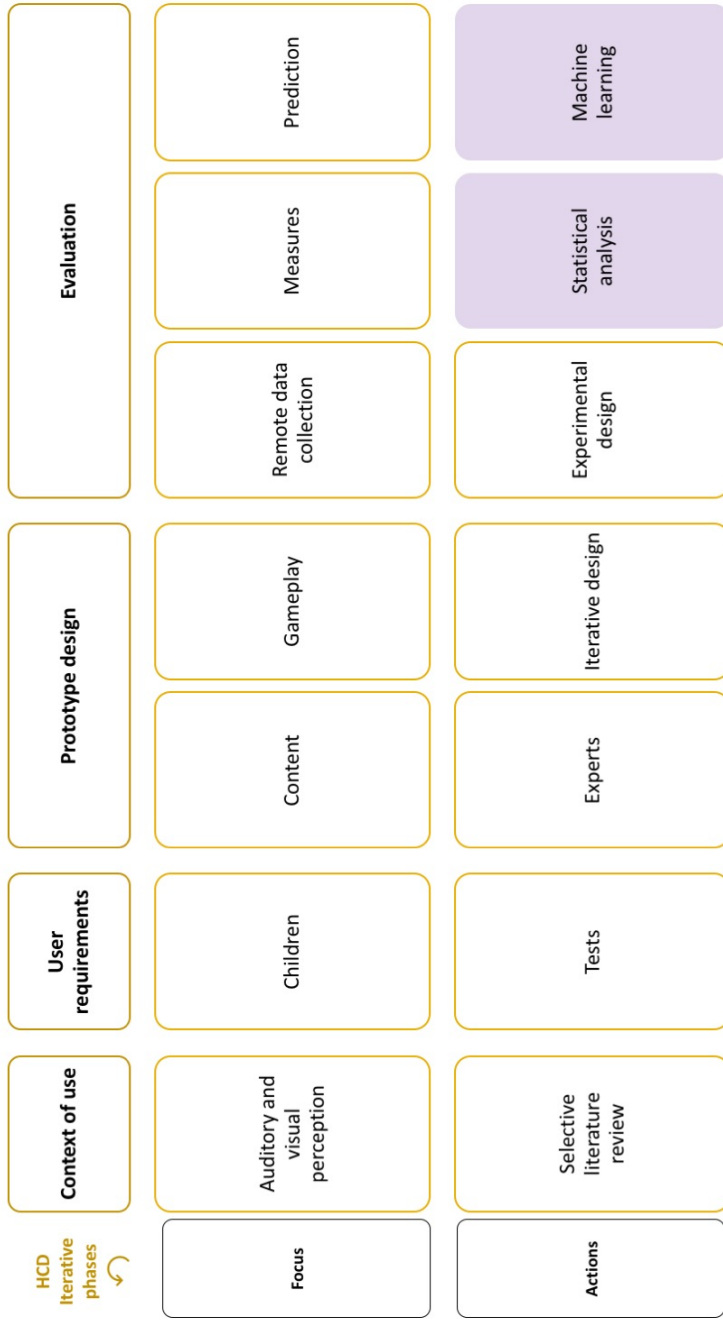


Figure 3.2: Overview of our actions for the human-centred design.

The user requirements and context of use define the content and gameplay for the prototypes, which are iteratively designed with the knowledge of experts. Furthermore, the HCD enhances the design, usability, and user experience of our prototype by avoiding external factors which could unintentionally influence the results. In particular, the early and iterative testing of the prototypes helps to avoid unintended interactions from participants or their supervisors. Example iterations are internal feedback loops of human-computer interaction experts or user tests, (e.g., five-user test). For instance, we discovered that children touch multiple times quickly on a tablet to interact. Because of the web implementation technique we used, a double click on a web application generally *zooms* in on the application on a tablet, which was not intended. Therefore, we controlled the layout setting for mobile devices to avoid the *zoom*-effect on tablets, which caused interruptions during the game. The evaluation requires the collection of remote data with the experimental design in order to use the dependent measures for statistical analysis and prediction with machine learning classifiers.

When taking into account participants with learning disorder, in our case participants with dyslexia, we need to address their needs in the design of the application and the experiment as well as consider the ethical aspects [10]. As dyslexia is connected to nine genetic markers and reading ability is highly hereditary [27], we support readability for participants' supervisors (who could be parents) with a large font size (minimum 18 points) [124].

We explain our activities for the DSRM and HCD phases used for this thesis in the next sections.

### 3.3 Problem Identification

As explained in Chapter 2, current approaches for detecting dyslexia require linguistic skills, expensive personnel, and/or spe-



cial hardware. The related work regarding screening for risk of dyslexia in pre-readers focuses mainly on gameplay. Only Lexa [98] published a prediction accuracy, but it focused on phonological features in a small data set ( $n = 56$ ) without precaution for overfitting. Additionally, children require general linguistic knowledge and phonological awareness to use it. To give children with dyslexia more time to practice, help them to avoid frustration, and increase their chances of success, a tool for early, language-independent detection is needed.

*Dyslexia is normally detected using linguistic knowledge. That is only possible when children have already developed reading skills. Is it possible to detect a child at risk of having dyslexia without linguistic knowledge?*

The next section describes the concept and research question developed to solve the identified problem.

### 3.4 Concept

From the selected literature review (presented in Chapter 2) and the identified problem (presented in the previous section), we know that dyslexia does not develop when one is learning to read and write. Rather, dyslexia already exists before learning to read and write. In fact, dyslexia has been associated with nine genetic markers, and reading ability is highly hereditary [27].

Common detecting and screening tools are mainly useful for children between the ages of 7 to 12 with linguistic knowledge. Evidence from the literature provides proof that there are language-independent indicators related to dyslexia, such as auditory working memory [82] or visual-spatial attention [156]. However, predicting dyslexia early on is still a challenge [6], which we address with language-independent content.

The related work focused on using one piece of evidence, *e.g.*, local visual search or rhythm. We combine findings from previous literature, which use visual and auditory perception to distinguish children with dyslexia from those without. This content is used to design a game environment that illuminates solid differences for predicting dyslexia in the future. At the same time, the game should be fun and not too difficult. We expect the people with dyslexia to make more mistakes and take more time than the control group. We advance previous approaches for screening risk of dyslexia by not focusing on linguistic knowledge and by using the same game content for every language. Using the same game content reduces the effort and time required to design different content for different languages, but more importantly, allows the content to be used for pre-readers.

At this point, a long-term study with pre-readers would be very time-consuming, since the effort to find participants is high, participants are less likely to be diagnosed, and a lot of time passes before results are available. An online study with readers has the advantage of reducing the effort and time required to design content, conduct various experiments for optimization, and increase the number of participants. Nevertheless, the language-independent content can be used for the screening of pre-readers who do not yet have any language skills.

### 3.4.1 Objectives

Based on the selected literature review (presented in Chapter 2) and the identified problem (previous section), we aim to answer with this thesis the following research questions:

**R1** Are there significant statistical differences between children with and without dyslexia when playing a game with auditory and visual content?

**R2** Is it possible to predict risk of dyslexia based on language-independent auditory and visual content using a game and machine learning for different languages?

**R3** Is it possible to predict risk of dyslexia based on generic language-independent visual and auditory content with various indicators using a game and machine learning?

### 3.4.2 Participant Requirements and Context

We conducted experiments with participants in the age range of 7 to 12, along with their supervisors (e.g., parent/legal guardian/teacher/therapist). We are familiar with the effects of dyslexia on individuals due to our close contact with diagnosed children, parents of children with dyslexia, teachers, and therapists, as well as because of the author's history with dyslexia. The extensive testing for and late detection of dyslexia often lead to harmful side effects [138]. These negative side effects are the reason we aim to design our experiments to be fun as well as motivating and engaging for children.

It is well known that children have more difficulties in maintaining attention over longer periods of time. Additionally, people with dyslexia should not be involved in long sessions or with many topics, as this could overwhelm or exhaust them [10]. Therefore, our games have limited numbers of rounds, and each game takes less than 15 minutes to play.

To gather more data, we set up an online experiment in which children can participate from different places, e.g., school, therapy, or home. For example, subjects will participate from different locations and environments, such as *Spain vs. Germany* or *school vs. home*. As a consequence, the technology used (such as headphones or device) will differ.

### 3.4.3 Integrating Gamification

Computer games are suitable for engaging children with dyslexia in reading and writing tasks [44, 75, 128]. Moreover, the quality of the game design could have an effect on the quality of the tests [153], which we test with the HCD. Like DYSL-X [45], we create a game to meet the actual goal (screening for dyslexia). To design a game in a non-game context, we used game elements to *gamify* our tasks for the online experiments. But because of the diversity of game elements used in applications as described in Appendix A.1, no reliable standard can be given to design gamified environments with game elements. Therefore, we analyzed the game elements by their quantity. A summary of *which* and *how* game elements have been applied in the different applications and frameworks is described in Appendix A.1. As gamification is a relatively young field of research, future work is necessary to give a comprehensive assessment of the topic, but this is not the main scope of this thesis.

The results of our systematic literature review analysis show that game elements are used heterogeneously, and only those game elements related to *dynamics*, *emotions* and *progression* are preferred in educational environments. The frequency distribution indicates preferences (see Appendix A.1, Figure A.1) for the different game dynamics and mechanics. Most game elements used for the environments are in the dynamics: *emotions* and *progression*.

Consequently, we use for our game design the game mechanics related to rewards (points), feedback (instant feedback), or challenges (time limit). Furthermore, our results confirm the acceptance of the term gamification for educational environments. But, since there is a widespread variety of game elements, the only common ground seems to be the definition of the term gamification made by Deterding [28]: Gamification applies game design elements in non-game contexts.

## 3.5 Content Design

The content we design is related to the evidence from the literature for distinguishing a person with dyslexia from one without. We design visual and auditory cues, as the reported evidence is connected to visual and auditory perception. We take into account the requirements of the participants as well as the fact that the participants' technological devices, such as headphones, will differ. Because different technological devices are used, we do not use promising indicators such as *intensity*, as variation among headphones and volume levels could easily influence perception of auditory cues.

We design the games with content that is strongly related to dyslexia. Furthermore, the gameplay is used to increase, for example, the cognitive load with time pressure, *i.e.*, *a limited amount of time is allotted for each round of the visual game*. However, at the same time, participants should not be overwhelmed by the content, which can lead to a loss of confidence or stop participation [10]. An iterative and user-focused approach like HCD helps to detect and avoid these kinds of influences. We describe the details of the designed content for each game in the section on game design.

## 3.6 Demonstration

We evaluate our concept and designed games with a data-driven approach, which we start by collecting data with the experimental design (see Table 3.1).

Initially, we designed the web-based prototype *MusVis* (mainly for desktop computers) with language-independent visual and auditory cues to investigate the cues across languages. Consequently, we conducted studies with German, Spanish and English students from 7 to 12 years old ( $n=178$  and  $n=313$ , see Chapter 4). The find-

Activity	Data collection	Pre-processing	Analysis	
Approach	Experimental design	Class labeling, data cleaning, dealing with missing values	Statistical analysis	Supervised machine learning classification
Tools and programming languages	Web Programming, PHP, JavaScript, jQuery and SQL-database	RStudio, Python and JupiterLab		

Table 3.1: Overview of the data-driven approach.

ings and experiences from our previous iteration of the approach were used to redesign the study and the content (see Chapter 5).

After we collected a sufficient amount of data, we pre-processed our data with data cleaning and decided how to deal with missing values. Subsequently, we analyzed our data with traditional statistical analysis to uncover differences between dependent measures and predict dyslexia using a machine learning classifier.

To achieve our thesis goals, we used different tools (see Table 3.1). Our games for the data collection are built with web programming tools and techniques such as *PHP*, *JavaScript*, *jQuery*, and a *SQL-database*. The pre-processing and analysis are mainly done with RStudio, Python, and JupiterLab. Next, we explain the general idea for our experimental design, data collection, and analysis.

### 3.6.1 Experimental Design

We base our experimental design on the following assumptions: (1) dyslexia does not develop when children come to school, but is already there before, (2) non-linguistic indicators can represent the difficulties a child with dyslexia has with writing and reading, and (3) dyslexia can be measured through the interaction data.

In our case, we need a certain age range to make sure a person with dyslexia has already been diagnosed (an official diagnosis

from an authorized specialist or a medical doctor) but has not yet been fully treated. Dyslexia is diagnosed in different languages, but always using spelling and reading tests connected to the language. To gather the data of this thesis, we had participants already diagnosed with dyslexia, to reduce the time to find the correct approach and increase the chances of designing a better game before having smaller children participate in a long-term study. Subsequently, we conducted different language-independent remote online experimental study to collect data to find differences in dependent variables between participants with and without dyslexia. Since the collection of participant data is costly in terms of time and resources, we evaluate our dependent variables with the pilot study before continuing the data collection.

We use the *Latin Squares* to counterbalance our conditions and avoid external factors, which means unintended influences on the participants, and therefore the analysis of the data is more difficult. The reason is that we save our data by the order of the Latin Squares, but analyze our data by the type of cue to avoid *order effects*. Since we used JupiterLab and Python for the machine learning prediction, we calculated the effect size with Cohens  $d$  [37], as it was more practical using the existing function for Python from [17].

### 3.6.2 Ethics

When conducting user experiments with participants who have difficulties, “*there may be a tension between what the community regards as being a rigorous methodology against what researchers can do ethically with their users*” [10]. Therefore, our research is in accordance with the ethical standards of the University and we address participants’ needs with the human-centred design (HCD) [67]. The *European Union* uses the *General Data Protection Regulation* to define how the data of a person, in regard to the processing of personal data, is protected [33]. The Ministry of Education, Science and Culture of Schleswig-Holstein (*Ministerium für Bildung,*

*Wissenschaft und Kultur, MBWK*) is in compliance with the GDPR [12, 33]. Therefore, we used the same approach and approval for the collection of data in Spain and in Germany.

The data collection for this work has been approved by the Ministry of Education, Science and Culture of Schleswig-Holstein (*Ministerium für Bildung, Wissenschaft und Kultur, MBWK*) and by the *Education Authority* of the State of Lower Saxony (*Niedersächsische Landesschulbehörde*) [88]. The State of Lower Saxony requires this additional permit for user studies at each school, and requires that no schools, teachers, or pupils are named. Specifics of the procedure are described in each Section for the study and the permits are in Appendix A.3.

### 3.6.3 Data Collection

The games and the user studies are designed with the *human-centred design* (HCD) framework [67] to collect the data set. It is relevant to include the HCD since collecting data related to e-health is challenging because of privacy and trust issues [5, 34].

Collecting data is costly in terms of time consumption and privacy issues, especially if the data is health-related. Therefore, we must make the most of our limited resources [5, 34]. Also, aggregated small data is sometimes better than individual level big data, and it also has the potential to reduce variations and privacy concerns [34].

In our case, small data means controlling the time and cost to collect data as well as providing reliable data for the analysis. That is why our final data set only includes participants that either have an official diagnosis or show no sign of dyslexia. We rely solely on self-reporting of diagnosis. However, in order to avoid the risk of fraudulent diagnosis reports by people wanting to participate for money, we exclusively use volunteer participants. An online screening test as a control test would ensure the quality of the data. However, since participation is voluntary and participants do



not have unlimited free time, this would further reduce the number of participants. Additionally, implementing or accessing such a test would take even more resources and time. The data sets we collected in our online experiment can help to achieve more realistic results [34].

We aim to protect the randomness and representation for our sample, but at the same time, want a precise data set according to the requirements of the *experimental study design*. However, with around 5% to 15% of the world population having dyslexia [2], we will handle an imbalanced data set, which we will address for the analysis. In our case, we need a certain age range to make sure a person with dyslexia has already been diagnosed but has not yet been fully treated. Dyslexia is diagnosed in different languages using different tests, e.g., spelling and reading tests, that are connected to that specific language. To investigate the possibility of a language-independent approach for both pre-readers and readers of different languages, we start with three languages with the same spelling error types [118, 123]: Spanish, German, and English. As explained in Chapter 2, the error words of children with dyslexia are phonetically and visually similar.

For online studies, the challenge is to engage participants over the Web, and in this case parents of children 7 to 12 years old. Therefore, we also recruited at schools and learning centers. Hence, parents and teachers had to trust us and take the time to participate in the online experiments that we conducted from 2017 to 2019.

### 3.7 Evaluation

The dependent measures are used to find differences between variables [37]. Second, the features are used as input for the classifiers to recognize patterns [13].

For our game content, we used language-independent indicators that have been related to dyslexia in lab studies. The specifics are explained in the content design sections of each game. It is challenging to design indicators so that measures can find differences between participant groups.

Although error rates are a common measure for dyslexia when reading, writing [51], letter naming [148], or even picture naming [71] is involved, error rates are not informative for dyslexia prediction in games. Games such as *Dydetective* [129] have shown that participants' mistakes are not informative for the machine learning prediction. Therefore, we used new dependent variables such as *click intervals*.

The comorbidity of dyslexia with other neuro-development disorders causes additive working memory deficits (as already explained in Chapter 2). The comorbidity makes the investigation of dyslexia more difficult [10]. Hence, we added questions regarding possible related diagnoses of the participants for *DGames* (e.g., ADHD diagnosis, hearing limitations).

The pre-processing (*class labeling, data cleaning, dealing with missing values*) of the data is done before the data analysis. For example, visual and auditory game data is merged from different CSV files to one participant's anonymous ID and class labels are renamed (Yes, diagnosed = 1; No, not diagnosed = 0). We excluded participants if the following values were not reported in the background questionnaire: age, dyslexia status, or gender. Parental consent was required before a child was allowed to participate in these studies. On one hand, we use the risk of dyslexia to include or exclude participants, and on the other hand, the status of dyslexia is meant to be determined using machine learning techniques.

Predictive traditional methods such as *regression* were designed before big data existed. Since we collected rather small data, it would be obvious to use these traditional methods. However, we did not use more traditional methods such as regression or approaches on step-wise regression since our data has a high vari-

ance which causes a high R-squared error and *multiple-collinearity* (*i.e.*, two or more variables have a high correlation). As dyslexia's origin is not fully decoded yet, more than one cause is assumed and therefore more than one indicator is stated for dyslexia, hence regression is not useful for the prediction. The data has complex dependencies, and using causal dependencies (*e.g.*, *having more correct or incorrect answers*), does not give a precise prediction.

Machine learning (ML) can find patterns in complex dependencies with statistical approaches to create predictive models. These methods have been proven to be more effective than traditional methods for modeling complex computational models. We use existing machine learning classifiers such as Random Forest with and without class weights, Extra Trees, and Gradient Boosting from the *Scikit-learn* library version 0.21.2 [143] for the prediction of dyslexia with language-independent content and small data sets.

A further advantage of ML is the extensive amount of techniques, which allow predictions for various use cases without being explicitly programmed for the use case. Our reported results focus on the class of people with dyslexia, *e.g.*, F1-score, accuracy, recall, and precision.

Since our collected data are considered *small data* [5, 34], we need to analyze them accordingly, *i.e.*, avoid over-fitting by using cross-validation instead of training, test, and validation sets [30]. When the data are small and sample variances high, we avoid over-fitting by using cross-validation instead of training, test, and validation sets, as well as by using classifiers' default parameters [30, 141, 154]. Because a small test or training set with high variances is not representative, prediction based on it could be misleading.

We address the danger of selecting the incorrect features [69] by taking into account knowledge from previous literature about the differences of children with and without dyslexia. For example, since there are two theories of the cause of dyslexia (visual vs. auditory [27]), we use subsets of visual and auditory features to explore their individual influences on the classifiers. We use subsets

of features (feature selection) to explore the influence on our data, while being aware of possible biases [154].

While we have small data, we do not optimize the input parameters of classifiers until we can hold out test data sets to evaluate changes as proposed by the *Scikit-learn documentation* [141]. We address the imbalanced data set for our binary classification problem (having dyslexia YES/NO) with different actions: using Random Forest with class weights (RFW) or computing the balanced accuracy.

Our aim is to detect a person with dyslexia. If we only consider the balanced accuracy, then we do not mainly focus on the detection of dyslexia, but rather on the overall accuracy of our model. Obtaining both high precision and high recall is unlikely, which is why we also report the F1-score (the weighted average between precision and recall) for dyslexia to compare our model's results.

We do not apply over-sampling to address our imbalanced data because the variances among people with dyslexia are broad; for example, difficulty levels or the individual causes for perception differences vary widely. We do not apply under-sampling to address our imbalanced data because our data set is already very small and under-sampling would reduce it to  $n < 100$ . The smaller the data set, the more likely it is to produce the unwanted over-fitting.

To avoid the risk of over-fitting our small data sets, we used 10-fold cross-validation and the default parameters suggested in the Scikit-learn library to avoid training a model by optimizing the parameters specifically for our data [143].

## 3.8 Outreach

On one hand, participants want to be informed about research results, and on the other hand, the interest in dyslexia research is immense. The participant calls in media and over social media have raised awareness of the topic of dyslexia. Around 32% of the par-

ticipants did not know or suspect that they might have dyslexia, and parents raised lots of questions in the communication.

Since participants with dyslexia or their parents (who may also have dyslexia) have difficulties with reading and writing, understanding research results might be an additional challenge for them [10]. However, reading research papers is not an everyday task, and even researchers are often overwhelmed by the number of existing research papers.

Clear and understandable communication for the different target groups and contexts about the results of the study is necessary to ensure the comprehension and visibility of the research results. Therefore, we provided results in different ways: newsletter ( $n = 400$ , accessed 09/July/2019), research social media group ( $n = 451$ , accessed 08/July/2019), personal website, Twitter/Instagram, and video stories throughout the years.

### 3.9 Discussion

This chapter presents the approach used for this thesis in terms of how methodologies are combined to screen children with language-independent content using a web game and a machine learning prediction using the interaction data. The advantage of using existing and well-known approaches such as the design science research methodology, human-centred design, and experiments design is the variety of methods that exist. However, describing and publishing results in this interdisciplinary environment can be challenging. It can be especially difficult to agree on terms that describe similar approaches or on similar terms that describe different approaches. We address this issue for this thesis by the integration of *design science research methodology* (DSRM) and *human-centred design* (HCD) (see Section 3.2).

Notably, the iterative evaluation and design of the prototype, first in Chapter 4 (with auditory and visual content that refers mainly to

one single acoustic or visual indicator) and then with the lessons learned in Chapter 5 (with auditory and visual content that refers to language-independent generic content related to various indicators), help to avoid unintended external factors which might influence the prediction results. The main challenge was to collect features that are meaningful for the prediction while also designing a game and content to achieve language-independent prediction using machine learning to answer our research questions.

Next, we present the first prediction with visual and auditory language-independent content using a game and a machine learning prediction using the interaction data.



# Screening Dyslexia with Auditory and Visual Cues

---

## 4.1 Introduction

Dyslexia is a *specific learning disorder* (prevalence 5-15% worldwide [2]) which affects the acquisition of reading and writing abilities. Children with dyslexia are often diagnosed with spelling and reading errors, sometimes after failing in school, even though dyslexia is not related to general intelligence. To be able to diagnose a person with dyslexia without using orthography or phonological awareness, new indicators, content and tools are needed.

In this chapter, we present how we designed a playful, easy and low-price web-based game and an approach for a possible earlier detection of dyslexia using machine learning models based in interaction data gathered from the language-independent game. We designed the game content taking into consideration the analysis of mistakes of people with dyslexia in different languages as well as other parameters related to dyslexia, such as auditory and visual perception.



Our main contribution is the first web-based game for screening risk of dyslexia on readers through a game that uses visual and auditory language-independent content and using a machine learning predictive model trained with interaction data. Our results show that the approach is feasible and that a higher prediction accuracy is obtained for German participants than for Spanish participants. The contributions of this chapter can be summarized as follows:

- The design of the web-based game and the language-independent auditory and visual content related to dyslexia (Section 4.3).
- The evaluation of our auditory game part with a usability test and the designed improvements for the web-based game with parents and children (Section 4.4).
- The experimental design setup, the statistical analysis and the predictive model provide results on how to design applications for children and parents, how participants with and without dyslexia differ from each other, and how the collected features can be used for the prediction (Sections 4.5, 4.6, 4.7, and 4.8).
- The discussion of how children with and without dyslexia can be distinguished (R1); for example, with seven significant measures based on language-independent auditory and visual cues in Spanish or for example, with one significant visual measure for all languages. Additionally, we discuss how auditory and visual language-independent content can be used for screening dyslexia in different languages (R2) with an accuracy of 0.74 and F1-score of 0.75 in German using a Random Forest and an accuracy of 0.69 and F1-score of 0.75 in Spanish using a Extra Trees (Section 4.9).

The research plan, the game design, the game content, and the analysis of this chapter were previously published in [106, 115, 117].

## 4.2 Methodology

Generally, we follow the methodology explained in Chapter 3 and point out here the important and case-specific steps for this application. Using the HCD and especially early user testing helps to avoid interaction mistakes (*i.e.*, *avoiding external factors for the user study*). Interaction mistakes might influence the prediction (as already explained in Chapters 2 and 3). With the usability first approach, we avoid influences for the iterative implementation of the new game (Section 4.4).

We focus our usability test on the new gameplay of the auditory part, the online consent, and the background questionnaire (called *DysMusic*). The reason is the visual gameplay is already tested with the application *Dytective* [127, 129]. After the usability changes are updated in the auditory part and integrated with the visual part, the application is called *MusVis*. Finally, we conducted a medium-scale remote online experimental study to collect the minimum sample size to explore a possible prediction with machine learning classifiers (see Section 4.7, 4.6, and 4.9).

## 4.3 Game Design

The aim of our web game, *MusVis*, (see Figure 4.1) is to measure the reactions of children with and without dyslexia while playing in order to find differences in the groups' behavior. A video of *MusVis* is available at <https://youtu.be/HeHERpYGA9Q>.

We designed the language-independent game content by taking into account knowledge of previous literature and selecting the most challenging content for children with dyslexia (CWD). Therefore, we designed language-independent content with auditory and visual cues. We describe in this section the game *MusVis*, which already integrates the changes suggested in Section 4.4.3 after the usability test. We designed an auditory part (see Figure 4.2) and

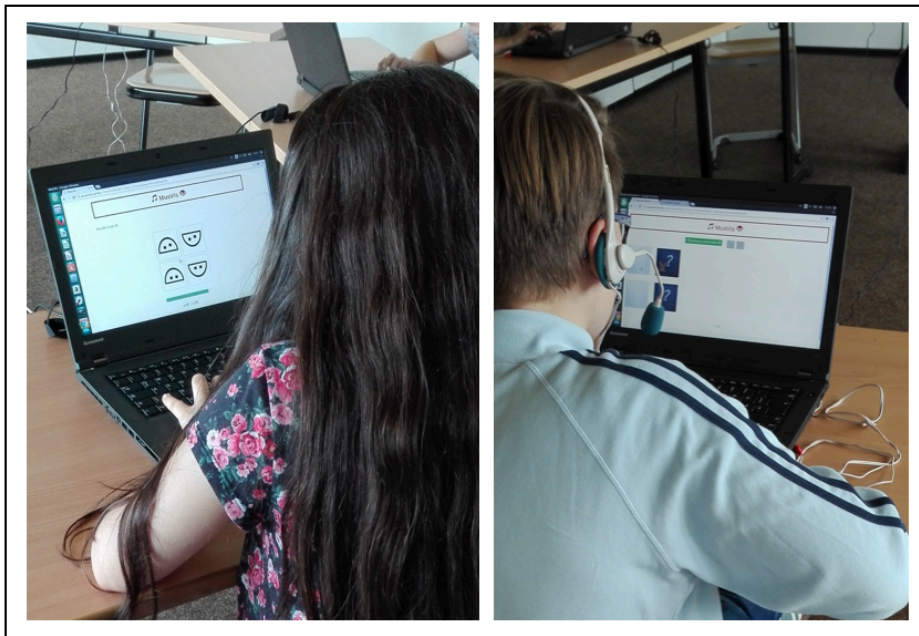


Figure 4.1: Participants playing the visual part (left) and the auditory part (right) of the Game *MusVis*. Photos included with permission.

a visual part (see Figure 4.3) of the game *MusVis* using features extracted from the literature. The game play for the auditory and visual parts is different due to unequal perception of auditory and visual cues, but both parts target general skills (e.g., short-term memory [47, 71, 92]), the phonological similarity effect [47], or the correlation of acoustic parameters speech [47, 165].

As is well known, children have more difficulty paying attention over a longer period of time. Therefore, the two games each have four stages, made up of eight rounds, each needing less than 10 minutes to play. Each part has four stages which are counter-balanced with *Latin Squares* [37]. Each stage has two rounds, which sums up to 16 rounds in total for the whole game. Each

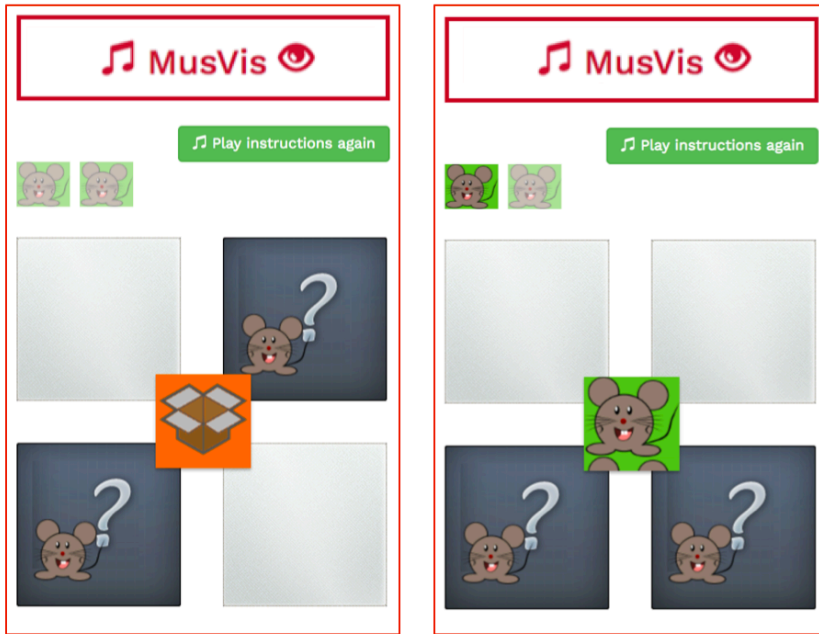


Figure 4.2: Example of the auditory part from the game *MusVis* for the first two clicks on two sound cards (left) and then a pair of equal sounds is found (right). The participant is asked to find two equal auditory cues by clicking on sound cards.

stage first has a round with four cards and then with six cards. We aim to address participants' motivation for both game parts with the design of the following game mechanics: rewards (points), feedback (instant feedback) or challenges (time limit), plus the game components (story for the game design). We identified these game mechanics as techniques to increase motivation through emotional engagement and visualize participants' progress (as explained in Chapter 3). The content design, user interface, interaction and implementation for the auditory and visual parts of the game are described in the following sections.

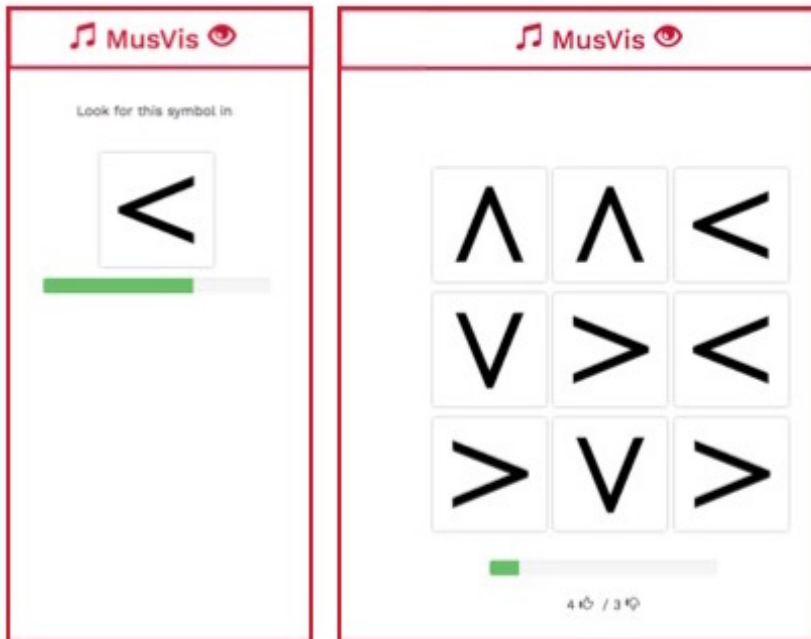


Figure 4.3: Example of the visual part of the game *MusVis* with the priming of the target cue *symbol* (left) and the nine-squared design including the distractors for each *symbol* (right).

#### 4.3.1 Auditory Cues

The auditory part modified the traditional game *Memory* in which pairs of identical cards (face down) must be identified by flipping them over [162]. We chose this game play because it is a well-known children game and could be easily transformed to use auditory cues. To create the auditory cues for the auditory part of our game *MusVis*, we used acoustic parameters; for example, to imitate the *prosodic* structure of language which is similar to the *phonological grammar* of music [99]. As explained in Chapter 2, musicians with dyslexia score better on auditory perception tests than the general population, but not on auditory working mem-

ory tests [82]. Auditory working memory helps a person to keep a sound in mind. We combined, for example, the deficits of children with dyslexia in auditory working memory with the results on the short duration of sounds [63] while taking the precaution of not measuring hearing ability [35]. We designed the auditory cues with an expert from the Music Technology Group from Universitat Pompeu Fabra and now we describe in detail our sound properties for each auditory cue.

To keep the game duration short, we only included very promising and easy to deploy acoustic parameters. Therefore, we used the acoustic parameters *frequency*, *length*, *rise time* and *rhythm* as auditory cues (see Table 4.1). Each auditory cue was assigned to a game stage (see Table 4.3), which we mapped to the attributes and literature references (see Table 4.2) that provide evidence for distinguishing a person with dyslexia.

For example, our *rhythm* stage uses the following characteristics: *complex vs. simple* [47, 63], *sound duration*, *rhythm* [63], *short-term memory* [47, 71], *phonological similarity effect* [47], and *correlation acoustic parameters speech* [47, 165].

The following is a short summary of the stages. Each acoustic stage has three auditory cues (we use MP3 for sound files). Only one acoustic parameter changes within a stage. Each stage is assigned to one acoustic parameter of sound, which is designed with knowledge of the analysis from previous literature (e.g., frequency or rhythm).

For the stage *frequency*, we use frequencies within the auditory perception range of a person, starting from 440 Hz.<sup>1</sup> We present the simple tone for a relatively short duration of 0.350s. Each auditory cue of this stage differs by 50 cents<sup>2</sup> intervals which is

---

<sup>1</sup>The 440 Hz is used for tuning instruments and is therefore in the auditory perception range of a person.

<sup>2</sup>Cent is a logarithmic unit of measure used for musical intervals. The twelve-tone octave (interval with double its frequency, e.g., C till C') is divided into 12 semitones of 100 cents each.

<b>Auditory cue</b>	<b>Sound Properties</b>
Always the same	<i>waveform</i> : sinus, mono files <i>amplitude</i> : 0.8 <i>auditory cues</i> : 2 to 3 sound files <i>frequency</i> : 440 Hz unless specified differently
Frequency ( <i>change of tone frequency</i> )	<i>2 cues</i> *: 440 Hz, 452.8929 Hz <i>3 cues</i> $\Delta$ : 2 previous sounds and 446.3998 Hz <i>fade in/out</i> : 0.025s <i>duration</i> : 0.350s
Length ( <i>change of tone length</i> )	<i>2 cues</i> : 0.350s, 0.437s <i>3 cues</i> : 2 previous sounds and 0.525s <i>fade in/out</i> : 0.025s
Rise Time ( <i>change in rise time</i> )	<i>2 cues</i> : 0.025s fade in, 0.250s fade in, both with fade out of 0.025s <i>3 cues</i> : 2 previous sounds and 0.025s fade in and 0.250s fade out <i>duration</i> : 0.500s
Rhythm ( <i>change in rise time for different auditory events</i> )	<i>2 cues</i> : (i) auditory events with rise time equal to 100ms, 100ms and 0.025s; and (ii) rise time equal to 100ms, 0.025s and 100ms <i>3 cues</i> : 2 previous sounds plus one with rise time equal to 0.025s, 100ms, 100ms) <i>duration</i> : 0.300s

Table 4.1: Auditory cues generated for the four tasks in *MusVis*. \* 2 cues: (1/2 semitone - 50 cents interval);  $\Delta$  3 cues: 3 sounds spaced by 25 cents (quarter of a semitone) - 2 previous ones.

<b>Key</b>	<b>Name</b>	<b>Description</b>
<b>CS</b>	Complex vs. simple	Children with dyslexia (CWD) recall significantly fewer items correctly in a lab study for long memory spans [47]. The rhythmic complexity did not have an effect on the difference between CWD and children without dyslexia (CC) [63].
<b>P</b>	Pitch	
<b>So</b>	Sound duration	Acoustic parameter differences in short tones (< 350 ms) are difficult to distinguish for a person with language difficulties [92].
<b>Ri</b>	Rise time	CWD or without showed significant differences when a comparing task used <i>rise time</i> [47]. Rise time and prosodic development are strongly connected and were shown to be most sensitive to dyslexia [63].
<b>Rh</b>	Rhythm	CWD show deficits in recalling the patterns of auditory cues [92]. However, rhythm modulations show no effect on the children performance [63].
<b>Sh</b>	Short-term memory	CWD show weaknesses in short-term memory tasks [92] when more items are presented [47]. Also, deficits can be frequently observed for the short-term auditory memory span [71].
<b>PSE</b>	Phonological similarity effect	CWD have difficulties with similar sounds and the <i>phonological neighborhood</i> when long memory spans are used [47].
<b>CAPS</b>	Correlation acoustic parameters speech	Since the <i>phonological grammar</i> of music is similar to the prosodic structure of language, music ( <i>i.e.</i> , a combination of acoustical parameters) can be used to imitate these features [165]. CWD are “ <i>reliably impaired in prosodic tasks</i> ” [47].

Table 4.2: Description of the auditory attributes which show promising relations to the prediction of dyslexia.



Attributes	Auditory					General		
	CS	P	So	Ri	Rh	Sh	PSE	CAPS
<b>Literature</b>								
Goswami et al.[47]	✓			✓		✓	✓	✓
Huss et al.[63]	✓	✓		✓	✓			
Johnson [71]						✓		
Overy [92]			✓		✓	✓		
Yuskaitis et al. [165]		✓						✓
<b>Stage</b>								
frequency	✓	✓	✓			✓	✓	✓
length			✓			✓	✓	✓
rise time	✓		✓	✓		✓	✓	✓
rhythm	✓		✓		✓	✓	✓	✓

Table 4.3: Mapping of the evidence from literature to distinguish a person with dyslexia, the attributes and general assumptions, and the stages of the auditory part of the game *MusVis*.

0.25 of a semitone (440 Hz to 452.8929 Hz to 446.3998 Hz). For the first round of two sound pairs, we use the 440 Hz and 446 Hz auditory cues.

In the stage *length*, each auditory cue has a different duration (0.350s, 0.437s, 0.525s), *i.e.*, tone length. The differences between the lengths of the auditory cues follow the suggested short duration (100ms) from Huss *et al.* [63]. For the first round of two sound pairs, we use the 0.350s and 0.525s auditory cues.

Each auditory cue of the stage *rise time* is designed with either a short fade in of 0.025s, a fade in of 0.250s, or a fade out of 0.250s. We use for the first round of two sound pairs the 0.025s and 0.250s fade ins.

The auditory cues of stage *rhythm* are designed with two intervals of rise time equal to 0.250s fade in and one interval equal to 0.025s fade in, in a changing order. The order of fade in for each auditory cue is changed according to the limit of possibilities (see example in Figure 4.4). We always use for the first round the

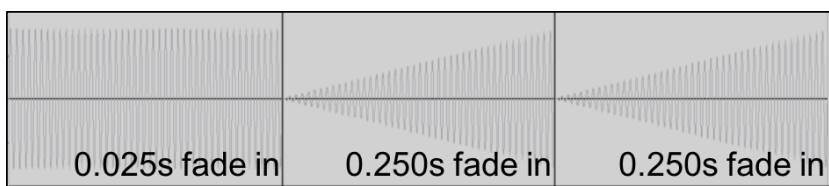


Figure 4.4: Waveform for the order of intervals for one auditory cue of the stage *Rise Time*. The example starts with a 0.025s fade in interval and then a 0.250s fade in interval followed by a 0.250s fade in interval.

two sound pairs with the order of rise time interval *0.025s*, *0.250s*, *0.250s* and the auditory cue with the rise time order reversed.

The auditory cues are generated with a simple sinus tone using the free software *Audacity*.<sup>3</sup> The exact parameters of each auditory cue are given in Table 4.1 and the auditory cues are available at *GitHub* [116]. Each stage has two rounds, with first two and then three auditory cues that must be assigned by choosing the same sound (see Figure 4.2). The arrangement of sounds (which auditory cue matches which card) is random for each round.

### 4.3.2 Visual Cues

The visual game play had a Whac-A-Mole interaction similar to the first round of *DyTECTIVE* [127]. But instead of using letter recognition as does *DyTECTIVE*, we used language-independent visual cues. An example for letter recognition would be finding the graphical representation of the letter /e/.<sup>4</sup>

We adapted the interaction design and content for this purpose (see Figure 4.3). For the visual game, we designed cues that have

<sup>3</sup>*Audacity* is available at <http://audacity.es/>, Last access: May 2019 .

<sup>4</sup>We used the standard linguistic conventions: '<>' for graphemes, '/ /' for phonemes and '[' ]' for phones.

the potential of making more cues with similar features and represent horizontal and vertical symmetries that are known to be difficult for a person with dyslexia in different languages [118, 126, 156].

To create the visual cues, we designed different visual representations similar to visual features of annotated error words from people with dyslexia [118, 126, 156] and designed the game as a simple search task, which does not require language acquisition. Each stage is assigned to a visual cue type (e.g., *rectangle or face*).

In the beginning, participants are shown the target visual cues (see Figure 4.3, left) for three seconds. They are asked to remember this visual cue. After that, the participants are presented with a setting where the target visual cue and distractors are displayed (see Figure 4.3, right). The participants try to click on the target visual cue as often as possible within a span of 15 seconds. The arrangement of the target and distractor cues randomly changes after every click.

The visual part has four stages, which are counter-balanced with *Latin Squares* [37]. Each stage is assigned to one visual type (*symbol, z, rectangle, face*) and four visual cues for each stage are presented. One visual cue is the target, which the participants need to find and click (see Figure 4.5, top). The other three visual cues are *distractors* for the participants. Each stage has two rounds (in total the number of rounds is 8) with first a 4-squared and then a 9-squared design (see Figure 4.3, right). The target and all three distractors are displayed in the 4-squared design. In the 9-squared design, the target is displayed twice as well as distractors two and three. Only distractor one is displayed three times. The stages from Figure 4.5 are summarized next.

**Stage *symbol*:** This stage uses two lines connected in an angle of less than  $30^\circ$  as the target visual cue and creates vertical symmetry. The distractor one is mirrored while distractors two and three are rotated by  $90^\circ$  and  $-90^\circ$ .

**Stage *z*:** The target visual cue for this stage is created with two lines parallel to each other connected with a diagonal line.

















	symbol	z	rectangle	face
target				
distractor 1				
distractor 2				
distractor 3				

Figure 4.5: Overview of the designed visual cues. The figure shows the target cue (top) and distractor cues (below) for the four different stages (*z*, *symbol*, *rectangle*, *face*) of the visual part of the game *MusVis*.

The diagonal line is drawn from the top right line end to the down left line end. This creates a horizontal symmetry of the visual cue. This representation looks very similar to the letter *z*, but we do not use phonological awareness of the letter (*i.e.*, the participants do not need to know that this is also a letter of the Latin alphabet). Distractor one is mirrored, while distractors two and three are rotated by 90° and -90°.

**Stage *rectangle*:** This stage is the shape of a square divided into two right-angled triangles, one of which is filled in. These shapes have by design vertical and horizontal symmetries, which we use to create a complex target. The 90° corner of one triangle

is placed in the top-right corner of the square and that of the other triangle in the bottom-left corner of the square. This creates a visual cue with different ways to perceive similarities within the cue. The distractors are rotated by 90°, 180° and 270°.

**Stage face:** This target visual cue has three visual cues combined (two symmetric dots placed horizontally inside a semicircle). The whole target cue is symmetrical on the vertical line. The target is rotated 180° for the first and third distractors. Additionally, the two dots are slightly staggered up and down for the second and third distractors.

### 4.3.3 User Interface

To avoid distractions or influences on the participants' behavior, the user interface is consistent in font size, as well as size, color, and shape of cards. To support readability for parents and supervisors, we used a large font size (minimum 18 points) [124]. The interactive elements (cards to be clicked within the game) are large enough to be clicked easily. The presentation of interactive elements (sound cards/squares) is the same within each game and does not differ in color or shape to avoid differences in perception.

### 4.3.4 Implementation

The game is implemented as a web application for the front-end with JavaScript, jQuery, CSS, and HTML5, and with a PHP server and a MySQL database for the back-end. One reason for this is access simplicity for remote online studies. Another reason is the advantage of adapting the application for different devices in future research studies.

## 4.4 Usability Test

First, internal feedback from HCI researchers improved the application and only minor changes on the game play needed to be done. After that, the usability test was conducted with children and parents who are not the authors of this thesis and are not familiarized with the research.

Since the auditory part is new, a user test [89] was conducted to discover (usability) problems which could have unintended influence on the planned study for predicting risk of having dyslexia. It should be mentioned that a five user test is a preliminary test for finding major usability problems and does not aim to find all usability problems.

### 4.4.1 Procedure

In a *within-subject design* [37], all participants played all four tasks of the game *DysMusic*. Only parents entered additional details for the study (e.g., background information) while using the *think aloud protocol* [16].

All participants played the web-based game *DysMusic* with the same tablet (Android Galaxy Tab A). The introduction video is available at <https://youtu.be/wIgcSMbE1VY>. Participants chose if they wanted to use headphones or not while playing (only one female parent used headphones). First, the parent read the study instructions and played with the sound cards while using the think aloud protocol. Afterwards, the parent or the first author filled in the background information for the child and the child played with the sound cards. After each sub-task, the first author asked the participant 'How difficult or easy was it to distinguish between the sounds?'. At the end of the game, each participant was asked if they had further comments on the interaction design of the game or the auditory cues.

## 4.4.2 Participants

We recruited ten participants: five children (users) and five parents. Two female parents (both age 35) and three male parents (ages 35, 40, and 40) participated. Each parent had two children, and five of their children took part in the user testing of *DysMusic*. Two female children (ages 3 and 8) and three male children (ages 5, 9 and 9) participated. All participants were German native speakers. Since there is no indication of significant differences in usability studies for people with or without dyslexia, we did not differentiate the two groups for the usability study.

## 4.4.3 Usability Improvements

We present now the results of the usability test and the changes we made to *DysMusic*. After these changes we call the game *MusVis*.

**Wording and Text:** Generally, the parents found the text easy to read and understand. They reported some spelling mistakes, which were mainly caused by the translation process from English to German, e.g., *study (English) vs. Studie (German)*. Parents also mentioned the large amount of text for *Online Consent*. One parent suggested only presenting the important information of the *Online Consent* and offering the possibility for further reading.

**Interaction:** All participants played the memory game with the auditory cues instantly and increased their speed after the first tasks, independently of the auditory cue. Only the youngest child (3 years old) had major problems with the amount of six sound cards for all auditory cues and did not find any sound pairs. We consider only using four cards when younger children play *DysMusic*. Some participants suggested including the button 'let's play' in the *game summary* in order to make the interaction more visible. The first author observed that the participants started to play faster, especially after the first task, and the delay of releasing the sound cards for

the next click helped to control the speed of the game interaction without being annoyingly slow.

**User Interface:** In general, all participants liked the structure, layout, and game cues (e.g., story), as well as the spoken motivational feedback: 'Yeah'. One participant commented that the footer of the game was very visible (large) and suggested making it smaller and less conspicuous. We did this change accordingly.

**Auditory cues:** All participants commented that they had to listen and concentrate carefully to be able to distinguish the auditory cues. Participants had different perceptions on how difficult it was to distinguish the sounds and find the card pairs depending on the auditory cues. But all participants always determined the first auditory cue of the first sub-task to be difficult, independent from the auditory cue (because of the counter-balanced design, the auditory cue order changed). This seems to be because it was the first time they played. For the second sub-task, they were already familiarized with the parameter and were able to name it. Two children and three parents mentioned difficulties in recognizing the auditory cue *Length*. One parent and one child of this group and another parent described more difficulties with the auditory cue *Frequency*. Only three parents reported difficulties in distinguishing the auditory cue *Rhythm* and two children mentioned difficulties for the auditory cue *Rise Time*.

**Functionality:** The motivational sounds between the exercises were not always played on the tablet and needed to be debugged for different devices. Besides, the video sometimes could not be played instantly, which may have been caused by a bad Wi-Fi connection. A change of the video player from *HTML 5: video-tag* to *YouTube: iframe-tag* prevented the loading problems.

**Other Comments:** In general, all participants found the task easy to understand. The children expressed more fun while playing than did the parents (e.g., by smiling or laughing). Three participants commented that the game was fun and all participants reacted positively to the spoken feedback 'Well done' when it was



played. We included more game sound cues (e.g., after each found pair we added a spoken feedback word like ‘Great’ or ‘Super’).

*DysMusic* is now called *MusVis* and incorporates the optimizations mentioned above and also the visual game part.

## 4.5 Experimental Design Setup

The usability test described in the previous section was used to improve the game *MusVis*, as already described in Section 4.3, to collect data with the experimental design. Here, we present the experimental design to collect data in an online experiment with our game *MusVis* to screen dyslexia [37].

We conducted a within-subject design study, which means that all participants played all game rounds [37]. As explained in Chapter 3, we first evaluate our indicators with a statistical comparison to find significant differences between the user groups (*i.e.*, between children with and without dyslexia). Subsequently, we collected more data to use predictive models to explore the prediction of dyslexia with our game and describe the participant groups according to the data sets. Only the game instructions were translated into each mother tongue.

The data collection for this user study has been approved by the Ministry of Education, Science and Culture of Schleswig-Holstein (*Ministerium für Bildung, Wissenschaft und Kultur*) and by the *Education Authority* of the state of Lower Saxony (*Niedersächsische Landesschulbehörde*), both in Germany.

### 4.5.1 Procedure

We follow the methodology description from Chapter 3. We summarize here the procedure and describe in more detail the steps. First, the parents were informed about the purpose of the voluntary study. Next, only after the parents had given their consent,

children were allowed to participate in the user study from home or from school, with me present or always available through digital communication. The communication with the participants was mostly via email or phone.

If the study was conducted in a school or learning center, the parental or legal guardian consent was obtained in advance, and the user study was supervised by the participant's supervisor (e.g., parent/legal guardian/teacher/therapist). After the online consent form was approved, we collected demographic data, which was completed by the participant's supervisor. This included the age of the participant, whether they had an existing dyslexia diagnosis (yes/no/maybe), and the native language. We asked the participant's supervisor to only set *YES* for a participant if the child had an official diagnosis from an authorized specialist or a medical doctor.

This was followed by explaining instructions for the user study to the participant's supervisor (e.g., turn up the volume, use headphones, play without interruptions, or explain and help your child only with the instructions of the games). Then a short video story for the auditory part was played. After that, every participant played first the auditory and then the visual part of *MusVis* (see Figure 4.1) and measurements were taken while playing.

At the end, two feedback questions were asked and the participant's supervisor could leave contact details to be informed about the results of the study.

Personal information of the participant's supervisor such as name or email is not published and is stored separately from the participant data for communicating results, if given. The name of the child is not collected and all data is stored on a password secured web server in Germany. Participants could choose not to participate or discontinue participation at any time during the study.

## 4.5.2 Participant Groups

The data includes only participants that completed all 16 rounds of the web game using a computer or a tablet. Dropouts happened mostly because participants used a different browser (e.g., *Internet Explorer* instead of *Google Chrome*) or a different device (tablet instead of a computer) for the statistical analysis of the pilot study.

Spanish participants diagnosed with dyslexia were mainly recruited over public social media calls from the non-profit organization *ChangeDyslexia* (<https://changedyslexia.org/>). We recruited German participants diagnosed with dyslexia mainly over support groups on social media. Also, some English speakers contacted us through this call, as our location (Barcelona, Spain) is very international. These participants played with the English instructions. The control groups for Spanish and German were recruited mostly with the collaboration of two Spanish schools and two German schools.

### **Participants of the Pilot Study**

For the statistical comparison ( $n = 178$ ), we included only participants that are either diagnosed with dyslexia ( $n = 67$ ) or do not have dyslexia ( $n = 111$ ). Thirteen participants were suspected of dyslexia and were therefore taken out of the analysis.

Each input method (*computer* vs. *tablet*) needs to be analyzed separately. We decided for the analysis of 178 to use a laptop or desktop computer for two reasons: (1) From prior game evaluation [44, 127], we know that readers are able to interact with the device, and (2) these devices are still more available than tablets [64].

We report separately the results for the Spanish participants ( $n = 108$ ), the German participants ( $n = 57$ ), and an analysis with all languages for the language-independent variables where we added English ( $n = 6$ ) and Catalan ( $n = 7$ ). The dependent variables that

show indications of the same tendency of results are considered to be *language-independent*.

For the analysis with all languages ( $n = 178$ ), we considered for the dyslexia group 67 participants that were diagnosed with dyslexia (33 female, 34 male). Their ages ranged from 7 to 12 years ( $\overline{age} = 9.8$ ,  $sd = 1.4$ ). For the control group, we considered 111 participants (67 female, 44 male). Their ages ranged from 7 to 12 years ( $\overline{age} = 10.5$ ,  $sd = 1.5$ ).

For the Spanish participants, we considered 41 participants diagnosed with dyslexia (23 female, 18 male). Their ages ranged from 7 to 12 years ( $\overline{age} = 9.5$ ,  $sd = 1.1$ ). For the control group, we took into account 67 participants (42 female, 25 male). Their ages ranged from 7 to 12 years ( $\overline{age} = 10.0$ ,  $sd = 1.2$ ).

For the German participants, we considered 17 participants diagnosed with dyslexia (5 female, 12 male). Their ages ranged from 7 to 12 years ( $\overline{age} = 10.7$ ,  $sd = 1.4$ ). For the control group, we had 40 participants (21 female, 19 male). Their ages ranged from 7 to 12 years ( $\overline{age} = 11.4$ ,  $sd = 1.4$ ).

## Participants of the Validation Experiments

For the predictive models, we took 313 participants into account, which include the ones from the pilot study. To have precise data, we took out participants that reported in the background questionnaire that they were suspected of having dyslexia but did not have a diagnosis ( $n = 48$ ).

The remaining participants were classified as diagnosed with dyslexia or not showing any signs of dyslexia (control group), as reported in the background questionnaire.

We separated our data into three data sets: one for the Spanish participants (ES,  $n = 153$ ), a second for the German participants (DE,  $n = 149$ ), and one for all languages (ALL,  $n = 313$ ) in which we included participants that spoke English ( $n = 11$ ). Participants

Data set	$n$	Dyslexia			Control				
		$n$	$\overline{age}$	female	male	$n$	$\overline{age}$	female	male
DE	149	59	10.22	21	38	90	9.58	42	48
ES	153	49	9.47	26	23	104	9.99	58	46
ALL	313	116	9.77	50	66	197	9.76	103	94

Table 4.4: Overview of the participants per data set for the validation experiments.

ranged in age from 7 to 12 years old. The data sets are described in Table 4.4.

The ALL data set ( $n = 313$ ,  $\overline{age} = 9.76$ ) contains 116 participants with dyslexia (50 female, 66 male,  $\overline{age} = 9.77$ ) and 197 as control (103 female, 94 male,  $\overline{age} = 9.76$ ). ES ( $n = 153$ ,  $\overline{age} = 9.82$ ) includes 49 participants with dyslexia (26 female, 23 male,  $\overline{age} = 9.47$ ) and 104 as control (58 female, 46 male,  $\overline{age} = 9.99$ ). DE ( $n = 149$ ,  $\overline{age} = 9.83$ ) is comprised of 59 participants with dyslexia (21 female, 38 male,  $\overline{age} = 10.22$ ) and 90 as control (42 female, 48 male,  $\overline{age} = 9.58$ ).

Participants played the game either in English, German or Spanish depending on their native language. We had some bilingual participants ( $n = 48$ ) in the Spanish data set (Spanish and Catalan) since the media call was done from a non-profit organization in the area of Catalonia (Spain). For these cases, we used the language they reported to be more comfortable with, which was used for the instructions of the game. We do not use the native language, but rather the language the game was played in as the criterion to split the data sets for three reasons.

First, the definition of a native language or mother tongue can be made easily when a participant speaks only one language. But this is not the case for bilingual participants because they might not be able to choose, and then we cannot distinguish the mother tongue or native language clearly [73]. Second, this question is a self-reported question and every participant’s supervisor might define it differently for each child. Finally, some bilingual speakers

spoke similar Latin languages (Spanish and Catalan). We consider these participants in the ES data set, as the instructions of the game were in Spanish.

### 4.5.3 Dependent Variables and Features

We conducted an online user study to collect the participants' responses (see participant features in Table 4.5) and the dependent variables. These variables were used for the statistical comparison of the pilot study and for the selection of the features. We used the following dependent variables for the statistical comparison:

#### **Auditory game part**

- *Duration round* (milliseconds) starts when round is initialized.
- *Duration interaction* (milliseconds) starts after the player clicks the first time on a card in each round.
- *Average click time* (milliseconds) is the duration of a round divided by the total number of clicks.
- *Time interval* (milliseconds) is the time needed for the second, third, fourth, fifth and sixth clicks.
- *Logic* we define it as *True* when in a round the first three clicked cards are different, otherwise, it is *False*.
- *Instructions* is the number of times the game instructions were listened by the player.

#### **Visual game part**

- *Number of hits* is the number of correct answers.
- *Number of misses* is the number of incorrect answers.
- *Efficiency* is the number of hits multiplied by the total number of clicks.

- *Accuracy* is the number of hits divided by the total number of clicks.

### **All part**

- *Time to the first click* (milliseconds) is the duration between the round start and the first user click.
- *Total number of clicks* is the number of clicks during a round.

We would like to further elaborate on the game measurement *Logic*, which is based on the direct experience of the user study. Some children may not have *really listened* to the sounds and played *logically*. As each round is designed such that the first two clicks never match, if the participant chooses for the third click a different card, s/he is increasing the chances of finding a match independent of the total amount of cards.

The descriptions of the participant features are in Table 4.5. Feature 1 was set with the language selected for the instructions. Features 2 to 8 were answered with the online questions by the participants' supervisor. Feature 9 was collected from the browser during the study experiment. The features for the data sets ALL, ES, and DE are the same. Each data set has 201 features per participant, where features 10 to 105 are the variables from the auditory part and features 106 to 201 are the variables from the visual part (Table 4.6).

<b>Participant features</b>	<b>Description</b>
<b>1</b> Age	It ranges from 7 to 12 years old.
<b>2</b> Gender	It is a binary feature either with <i>female</i> or <i>male</i> value.
<b>3</b> Language	It is either <i>Spanish, German or English</i> .
<b>4</b> Native Language	It indicates if the language used for the instructions is the first language of the participants, being <i>Yes, No or Maybe</i> .
<b>5</b> Instrument	It indicates if a participant plays a musical instrument, being <i>No, Yes, less than 6 months or Yes, over 6 months</i> .
<b>6</b> Memory	It indicates how well the participant knows the visual <i>Memory</i> game, being <i>Participant gave no answer, Participants does not know the game, Played once, Played a few times or Played a lot</i> .
<b>7</b> Rating Auditory Part	It indicates the self-reported answer with a 6-level <i>Likert scale</i> [37] to the statement: 'the auditory part was easy for the participants.' The values are <i>Answer unknown, Strongly disagree, Disagree, Undecided, Agree or Strongly Agree</i> .
<b>8</b> Rating Visual Part	It indicates the self-reported answer of the statement: 'the visual part was easy for the participants.' (We used the same <i>Likert scale</i> from feature 7.)
<b>9</b> Device	It is the device the participants used and is a binary feature with the <i>Computer or Tablet</i> value.

Table 4.5: Description of participant features.



<b>Auditory features</b>	<b>Visual features</b>
<b>10–17</b> Time to click.	<b>106–113</b> Time to click.
<b>18–25</b> Total clicks.	<b>114–121</b> Total clicks.
<b>26–33</b> Duration per round.	<b>122–129</b> Correct answers.
<b>34–41</b> Duration interaction.	<b>130–137</b> Wrong answers.
<b>42–49</b> Average click time.	<b>138–145</b> Accuracy.
<b>50–57</b> Logic.	<b>146–153</b> Efficiency.
<b>58–65</b> 2nd click interval.	<b>154–161</b> 2nd click interval.
<b>66–73</b> 3rd click interval.	<b>162–169</b> 3rd click interval.
<b>74–81</b> 4th click interval.	<b>170–177</b> 4th click interval.
<b>82–89</b> 5th click interval.	<b>178–185</b> 5th click interval.
<b>90–97</b> 6th click interval.	<b>186–193</b> 6th click interval.
<b>98–105</b> Instructions.	<b>194–201</b> Time last click.

Table 4.6: On the left are features 10 to 105 for the auditory part and on the right are features 106 to 201 for the visual part of the game MusVis.

## 4.6 Predictive Models Setup

In this section, we present the machine learning approach for the data sets ALL ( $n=313$ ), ES ( $n=153$ ), and DE ( $n=149$ ). First we explain the choice of predictive models and then the choice of feature selection.

### 4.6.1 Model Selection

We used Random Forest (RF), Random Forest with class weights (RFW), Extra Trees (ETC), Gradient Boosting (GB), and the Dummy Classifier (Baseline), which are described in the Scikit-learn version 0.21.2 [143]. As already explained in Chapter 3, we address the risk of over-fitting our small data sets with 10-fold cross-validation and the default parameters suggested in the Scikit-learn library to avoid training a model by optimizing the parameters specifically for our data [143].

To explore the best prediction conditions we used the feature selection as described in the next section.

### 4.6.2 Feature Selection

We rank the most informative features with *Extra Trees*. The results show a flat distribution for all three data sets and a step at the information score of 0.008: ALL ( $n=33$  features), ES ( $n=41$  features), and DE ( $n=38$  features). The comparison of the most informative features reveals that the data sets have only a few features in common, e.g., four features for Spanish and German (Logic, 6th click interval, total click, duration interaction) or only 16 features in ALL compared to Spanish and German. Visual and auditory features are equally represented in the ranking of the most informative features; for example, ALL has 16 auditory features and 14 visual features.

The biggest step in the informative ranking for all three data sets is between the fifth and sixth informative features, e.g., for ALL the step is between the visual part (cue *Z*, 4 cards) *Efficiency* with the informative score of 0.0128 and the auditory part (cue *Rhythm*, 6 cards) *Time 5th click* with a score of 0.0104. The only dependent variables with the same tendency are *Number of misses* and *Total clicks* from the visual game part, but the features from the different rounds for the different data sets are mainly not under the 33 informative features (ALL 2/16, ES 3/16 and DE 6/16).

We explore the influence of the features on the accuracy of the different selected feature subsets (i.e., 201, 33, 27, 20 and 5). We choose the subset 201 since it contains all possible features and the subset 5 since it contains the most informative features. Additionally, we choose the first 33 features because they represent the next step of the most informative features. Finally, we choose 20 features randomly as a comparison to 33 features. Finally, we selected only the 27 features that have the same tendency and have been answered by the participant's supervisor because they are mainly not under the most informative feature ranking.

## 4.7 Statistical Analysis

We present here the results for the *statistical comparison* of dependent measurements to find differences between children with and without dyslexia for different languages and for the *predictive models* to report the possibility of predicting the risk of having dyslexia. In order to find out whether we have new indicators to predict people with dyslexia after playing *MusVis*, we analyzed the dependent variables from our independent within-subject study for the three groups: *Spanish (ES)*, *German (DE)*, and *all languages (ALL)*.

The pilot study collected data from 178 participants (which were later included into our validation data set of  $n = 313$ ) to find significant differences on the game measurements. We followed the

same steps of the pilot study ( $n = 178$ ) for the validation data set ( $n = 313$ ) to compare the findings. Therefore, we apply first the *Shapiro-Wilk test* and then the *Mann-Whitney U Test* since all game measures ( $n = 54$ ) are not normally distributed. We use the Bonferroni correction ( $p < 0.00$ ) to avoid type I errors [37].

### 4.7.1 Pilot-Study

We present the results of the pilot-study ( $n = 178$ ) separated by data set for German (see Table 4.7), for Spanish (see Table 4.8), and for all languages (see Table 4.9).

The dependent variables are categorized for Spanish and German according to the tendency that participants with dyslexia had compared to the control group within each language (see Table 4.10). An example of a language-independent variable is the dependent variable *hits*, because the dyslexia group has in German and Spanish significantly fewer correct clicks (Spanish 5.7; German 5.6) than the control group (Spanish 6.6; German 6.3). The dependent variable *duration* is an example of the opposite trend because the dyslexia group for Spanish takes significantly more time while the German participants with dyslexia take less time compared to their language's control group. The dependent variables were included in the overview of all languages only if the tendency was similar between groups (Table 4.9). We consider the variables in Table 4.10 as a first step to providing evidence for language-independent detection of dyslexia.

We use the effect size (see Section 3) in Tables 4.7 and 4.8 only for the significant results. First, we report the results for the auditory part and then for the visual part of the game.

**Total number of clicks (auditory)** is not language-independent. The tendency of results is opposite between participants with dyslexia (Spanish,  $m = 11.3$  and German,  $m = 10.6$ ) compared to participants without dyslexia (Spanish,  $m = 11.0$  and German,  $m = 10.8$ ). This means that German par-

Dependent variables <i>DE</i> ( $n = 57$ )	Control		Dyslexia		Mann-Whitney U			
	mean	sd	mean	sd	W	p-value	z	effect size
<b>Auditory</b>								
Total clicks	10.8	5.4	10.6	4.4	20880	0.49	-0.70	0.09
4th click	1.8s	0.8s	2.0s	1.2s	19218	0.05	-2.00	0.26
6th click	1.7s	0.7s	1.6s	0.7s	21580	0.89	-0.14	0.02
Duration	28.5s	16.9s	27.9s	13.0s	20542	0.34	-0.95	0.13
Average click time	2.6s	0.8s	2.6s	0.5	19.708	0.11	-1.59	0.21
<b>Visual</b>								
Total clicks	7.2	3.1	6.8	2.7	23887	0.10	1.67	0.22
Time to first click	2.4s	1.5s	2.5s	1.1s	19314	0.06	-1.90	0.25
Hits	6.3	2.8	5.6	2.6	24675	0.02	2.28	0.30
Misses	0.9	2.1	1.2	2.3	20718	0.36	-0.92	0.12
Accuracy	0.60	0.49	0.59	0.49	22084	0.83	0.30	0.04
Efficiency	2.8s	2.3s	3.2s	2.9s	19357	0.06	-1.87	0.25

Table 4.7: Overview of all selected dependent variables for the auditory and visual parts of the game *MusVis* for German.

Participants with dyslexia click less often compared to the German control group, while Spanish participants with dyslexia click more often compared to the Spanish control group. The *total number of clicks* did not reveal significant differences in *total clicks* for Spanish ( $W = 86231$ ,  $p = 0.63$ ,  $r = 0.05$ ) or German ( $W = 20880$ ,  $p = 0.49$ ,  $r = 0.09$ ). The effect size for Spanish and German is nearly zero, so it is considered to have no effect [37].

**Click time interval (auditory)** is not language-independent over all click intervals. Therefore, we do not report any click intervals for all languages. Hence, participants with dyslexia (Spanish 4th click interval  $m = 2.0s$ , German 4th click interval  $m = 2.0s$  and Spanish 6th click interval  $m = 1.7s$ ) take more time before they make the next click than the control group (Spanish 4th click interval  $m = 1.6s$  and German 4th click interval  $m = 1.8s$  and 6th click

Dependent variables <i>ES</i> ( $n = 108$ )	Control		Dyslexia		Mann-Whitney U			
	mean	sd	mean	sd	W	p-value	z	effect size
<b>Auditory</b>								
Total clicks	11.0	5.5	11.3	6.0	86231	0.63	-0.48	0.05
4th click	1.6s	0.7s	2.0s	1.3s	63658	<b>7e-12</b>	-6.80	0.66
6th click	1.5s	0.8s	1.7s	1.2s	76762	<b>2e-3</b>	-3.13	0.30
Duration	27.5s	17.1s	34.3s	27.0s	72316	<b>1e-5</b>	-4.38	0.42
Average click time	2.5s	0.8s	3.0s	1.2s	59028	<b>2e-16</b>	-8.11	0.78
<b>Visual</b>								
Total clicks	8.0	3.3	6.7	2.7	110000	<b>3e-10</b>	6.25	0.60
Time to first click	2.3s	1.4s	2.7s	1.8s	75566	<b>5e-4</b>	-3.47	0.33
Hits	6.6	2.9	5.7	3.0	105670	<b>5e-7</b>	5.02	0.48
Misses	1.3	3.1	1.0	1.8	86340	0.62	-0.50	0.05
Accuracy	0.60	0.50	0.57	0.50	90432	0.43	0.83	0.08
Efficiency	2.8s	2.6s	3.1s	2.8s	73301	<b>4e-5</b>	-4.10	0.39

Table 4.8: Overview of all selected dependent variables for the auditory and visual part of the game *MusVis* for Spanish.

Dependent variables ALL ( $n = 178$ )	Control		Dyslexia		Mann-Whitney U	
	mean	sd	mean	sd	W	p-value
Total clicks	7.6	3.2	6.8	2.7	276120	<b>3e-7</b>
Time to first click	2.4s	1.5s	2.6s	1.6s	210850	<b>3e-4</b>
Hits	6.5	2.9	5.8	2.9	272180	<b>4e-6</b>
Accuracy	0.60	0.49	0.58	0.49	240780	0.66
Efficiency	2.8s	2.5s	3.1s	2.7s	209740	<b>2e-4</b>

Table 4.9: Overview of dependent variables with the same tendency, which are all from the visual part of the game for ALL.

	<b>Language-Independent</b>	
	<i>n</i> = 178	<i>n</i> = 313
<b>Auditory</b>		
Total clicks	No	No
4th click interval	No	No
6th click interval	No	No
Duration	No	No
Average click time	No	No
<b>Visual</b>		
Total clicks	<b>Yes</b>	<b>Yes</b>
Time to first click	<b>Yes</b>	No
Hits	<b>Yes</b>	No
Misses	No	<b>Yes</b>
Accuracy	<b>Yes</b>	No
Efficiency	<b>Yes</b>	No

Table 4.10: Overview of all dependent variables showing the language-independent results between the German and Spanish groups.

interval  $m = 1.5s$ ). But German participants with dyslexia (6th click interval  $m = 1.6s$ ) take less time before they make the next click than the German control group ( $m = 1.7s$ ). The *4th time interval* ( $W = 63658, p = 7e - 12, r = 0.66$ ) and the *6th click interval* ( $W = 76762, p = 2e - 3, r = 0.30$ ) are significant for Spanish but not for German ( $W = 21580, p = 0.89, r = 0.02$ ). The effect size for *4th click time interval* Spanish is considered to be large, while the effect sizes for *6th click time interval* for Spanish and the *4th click time interval* for German are considered to be medium. We report only the fourth and sixth click intervals for the auditory game part since the first three intervals do not show, as expected due to the game design, any significant differences between the groups.

**Duration (auditory)** is not language-independent. Hence, Spanish participants with dyslexia ( $m = 34.3s$ ) take more time to

find all pairs and finish the round than the Spanish control group ( $m = 27.5s$ ). But German participants with dyslexia ( $m = 27.9s$ ) take less time before they find all pairs than the German control group ( $m = 28.5s$ ). The *duration* is significant for Spanish ( $W = 72316$ ,  $p = 1e - 5$ ,  $r = 0.42$ ) but not for German ( $W = 20542$ ,  $p = 0.34$ ,  $r = 0.13$ ). The effect size for Spanish is considered to be medium and for German it is considered small.

**Average click time (auditory)** is not language-independent. Spanish participants with dyslexia ( $m = 3.0s$ ) take, on average, more time per click than the Spanish control group participants ( $m = 2.5s$ ). But German participants with and without dyslexia take about the same time ( $m = 2.6s$ ). Spanish participants with dyslexia spend significantly more time for each click ( $W = 59028$ ,  $p = 2e - 16$ ,  $r = 0.78$ ), while we cannot measure a difference for the German participants ( $W = 19708$ ,  $p = 0.11$ ,  $r = 0.21$ ). The effect size for Spanish is considered to be large and for German it is small.

**Total number of clicks (visual)** is language-independent. Participants with dyslexia ( $m = 6.7$ ) clicked significantly fewer times than participants without dyslexia ( $m = 8.0$ ) for Spanish ( $W = 110000$ ,  $p = 3e - 10$ ,  $r = 0.60$ ). The effect size for Spanish is considered to be large. The German participants have the same trend for the control group ( $m = 7.2$ ) compared with the group of participants with dyslexia ( $m = 6.8$ ,  $W = 23887$ ,  $p = 0.10$ ,  $r = 0.22$ ). The effect size for German is considered small. Because the trend is the same for German and Spanish participants, we provide the *total number of clicks* for ALL, which confirms the significant difference ( $W = 276120$ ,  $p = 3e - 7$ ).

**Time to the first click (visual)** is language-independent. This means that participants with dyslexia (Spanish,  $m = 2.6s$ ) and German ( $m = 2.5s$ ) take more time before they make the first click than the control group (Spanish,  $m = 2.3s$  and German,  $m = 2.4s$ ). The *time to the first click* is significant for Spanish ( $W = 89450$ ,  $p = 1e - 3$ ,  $r = 0.30$ ) but not for German ( $W = 19314$ ,  $p = 0.06$ ,  $r = 0.25$ ). The effect sizes for Spanish and German are considered medium.



Because the trend is the same for both languages, even though it is only significant for Spanish, we provide the analysis for the *time to the first click* with a significant difference for all languages ( $W = 210850, p = 3e - 4$ ).

**Hits** is language-independent. Hence, participants with dyslexia (Spanish,  $m = 5.7s$  and German,  $m = 5.6s$ ) have fewer hits than the control group (Spanish,  $m = 6.6s$  and German,  $m = 6.3s$ ). *Hits* is significant for Spanish ( $W = 105670, p = 5e - 7, r = 0.48$ ) and for German ( $W = 24675, p = 0.02, r = 0.30$ ). The effect sizes for Spanish and for German are considered medium. Because the trend is the same for both languages, we provide the analysis for *hits* with a significant difference for all languages,  $W = 272180, p = 4e - 6$ .

**Misses** is not language-independent. Hence, Spanish participants with dyslexia ( $m = 1.0$ ) make fewer mistakes than the Spanish control group ( $m = 1.3$ ). But German participants with dyslexia ( $m = 1.2$ ) make more mistakes than the German control group ( $m = 0.9$ ). *Misses* has no significant difference for Spanish ( $W = 86340, p = 0.62, r = 0.05$ ) or German ( $W = 20718, p = 0.36, r = 0.12$ ). The effect size is considered small for both languages.

**Accuracy** is language-independent. There were no differences for participants with dyslexia (Spanish,  $m = 0.57$  and German,  $m = 0.59$ ) and the control group (Spanish,  $m = 0.60$  and German,  $m = 0.60$ ) in *accuracy*. *Accuracy* is not significantly different for Spanish ( $W = 90432, p = 0.43, r = 0.08$ ) or German ( $W = 22084, p = 0.83, r = 0.04$ ). Because the trend is the same for both languages, we report the *accuracy* for ALL, which has no significant difference for all languages ( $W = 240780, p = 0.66$ ).

**Efficiency** is language-independent. Hence, participants with dyslexia (Spanish,  $m = 3.1s$  and German,  $m = 3.2s$ ) take more time for a hit than the control group (Spanish,  $m = 2.8s$  and German,  $m = 2.8s$ ). *Efficiency* is significant for Spanish ( $W = 73301, p = 4e - 5, r = 0.39$ ) but not for German ( $W = 19357, p = 0.06, r = 0.25$ ). The effect size is considered medium for both languages. Because the trend is the same for both languages, we provide the

analysis for *efficiency* with a significant difference for all languages ( $W = 209740$ ,  $p = 2e - 4$ ).

Children and parents provided positive ( $n = 44$ ) and negative ( $n = 7$ ) feedback about the gameplay or content. The following are translated quotes from children that show examples of the positive feedback we received. A boy (8 years) who participated in a school said: “*This was so cool! It was the best day at school ever*”. A girl (12 years) wrote in the web feedback input field that “*it was fun and not boring!*” and a boy (10 years) wrote “*I love this game*”. The positive feedback was provided by all age groups. However, some provided negative feedback. A girl (12 years) said it was “*not exciting more boring,*” while a boy (12 years) wrote that the “*game started too fast*”.

#### 4.7.2 Validation

We present the results of the statistical analysis for the validation data ( $n = 313$ ) separated by language and for all languages (see Table 4.11). Additionally, we compare the statistical analysis results from the pilot-study ( $n = 178$ ) with the validation data set ( $n = 313$ ).

The ES data set ( $n = 153$ ) has seven dependent variables with significant differences between groups: *4th click interval*, *duration round*, *average click time*, *total number of clicks*, *time to the first click*, *number of hits*, and *efficiency*. The ES data set ( $n = 153$ ) confirmed the results of the pilot study ( $n = 178$ ). The visual accuracy is now also significant ( $p = 0.026$ , before  $p = 0.43$ ). All other game measurements decreased the significance by slightly increasing the p-value (visual efficiency from  $4e - 5$  to  $1e - 4$ ). The data set ES has 9 significant variables that distinguish a person with or without dyslexia.

Part	Lang.	DM	Control mean	sd	Dyslexia mean	d	Mann-Whitney U W	p-value	effect size
Visual	ALL	Total clicks	6.8	2.7	7.2	3.2	670194	<b>2e-04</b>	0.14
		Misses	1.2	2	1.3	2.7	713627	0.14	0.05
	ES	Total clicks	6.8	2.7	7.7	3	132207	<b>3e-08</b>	<b>0.31</b>
		First click	2.63s	1.69s	2.26s	1.22s	141938	<b>1e-04</b>	0.27
		Hits	5.8	3	6.5	2.9	136904	<b>2e-06</b>	0.25
		Misses	1	1.7	1.2	2.7	157086	0.12	0.07
		Accuracy	0.82	0.27	0.85	0.26	153012	0.03	0.10
		Efficiency	3.1s	2.6s	2.75	2.4s	142162	<b>1e-04</b>	0.14
	DE	Total clicks	6.7	2.6	6.8	3.3	169439	0.47	0.03
		First click	2.50s	1.32s	2.58s	1.56s	168932	0.43	0.06
		Hits	5.4	2.6	5.3	2.8	164224	0.16	0.05
		Misses	1.3	2.1	1.5	2.8	166140	0.24	0.09
		Accuracy	0.81	0.27	0.78	0.29	165688	0.22	0.08
Efficiency		3.2s	2.4s	3.5s	2.9s	167288	0.33	0.10	
Audit.	ES	Total clicks	11.3	6	10.9	5.5	157282	0.15	0.07
		<b>4th click</b>	2.0s	1.3s	1.7s	1.0s	131228	<b>1e-08</b>	0.29
		6th click	1.7s	1.1s	1.6s	0.9s	152772	0.04	0.15
		Duration	32.6s	69.9s	24.7s	18.2s	142726	<b>2e-04</b>	0.19
		Average	3.0s	2.7s	2.6s	0.9s	121966	<b>5e-13</b>	0.29
		Total clicks	11.1	5.5	11.5	6.6	166340	0.27	0.07
	DE	4th click	1.9s	1.0s	2.0s	1.0s	167184	0.32	0.01
		6th click	1.8s	0.8s	1.9s	1.3s	163076	0.12	0.12
		Duration	27.1s	18.6s	29.4s	22.9s	163994	0.15	0.11
		Average	2.7s	0.8s	2.8s	1.0s	166194	0.26	0.11

Table 4.11: Overview of dependent variables for visual (top) and auditory (below) features of DGames. Significant results are in bold.

For the data set ALL ( $n = 313$ ) we consider only dependent variables with the same tendency as for the pilot study ( $n = 178$ ). We categorize the tendency (e.g., *playing faster or having more clicks*) by the group (dyslexia compared to control group) *mean* of the dependent variables within the same language. ALL ( $n = 313$ ) has two visual game measurements (*number of misses* and *total clicks*) with the same tendency while the pilot study had five for the visual game (*total clicks, time to the first click, hits, accuracy, and efficiency*).

The DE data set ( $n = 149$ ) confirmed the results of the pilot study ( $n = 57$ ) with no significant dependent variables. The *means* of the dependent measurements for DE are all very close (e.g., the *time to the first click* is  $2.58s$  for the control group and  $2.50s$  for the dyslexia group).

We can confirm that misses did not reveal significant differences for German or Spanish, even though the tendency is now the same for both languages. On the other hand, the total number of clicks is still significant.

To sum up, we confirmed one significant dependent variable in ALL ( $n = 313$ ), seven significant dependent variables for ES ( $n = 153$ ), and no significant dependent variables for DE ( $n = 149$ ).

## 4.8 Prediction using Machine Learning

We processed our data sets with different classifiers and different subsets of features, following Section 4.6. We computed the *balanced accuracy* for our binary classification problem to deal with imbalanced data sets; for example, ALL data set has dyslexia 37% vs. control 63%. The best results obtained are presented in Table 4.12.

As described in the Section 4.6.2, the ranking of the informative features is different for the three data sets. Hence, we explore the influence of different subsets of features, namely: (1) all represented features (201 features); (2) the 5 most informative features; (3) the 33 most informative features, as this was the next natural

Clas.	Data set	Feat.	Recall	Precision	F1	Accuracy
<b>RF</b>	<b>DE</b>	<b>5</b>	0.77	0.78	<b>0.75</b>	<b>0.74</b>
RFW	DE	5	0.75	0.75	0.74	0.73
Baseline	DE		0.60	0.37	0.46	0.50
<b>ETC</b>	<b>ES</b>	<b>20</b>	0.76	0.76	<b>0.75</b>	<b>0.69</b>
RF	ES	5	0.74	0.73	0.72	0.65
Baseline	ES		0.68	0.46	0.55	0.50
<b>GB</b>	<b>ALL</b>	<b>20</b>	0.66	0.65	<b>0.65</b>	<b>0.61</b>
GB	ALL	5	0.64	0.64	0.63	0.59
Baseline	ALL		0.63	0.40	0.49	0.50

Table 4.12: Best results of the different classifiers, features and data sets. Results are ordered by the best F1-score and accuracy.

informative subset; (4) 20 random features selected from (3); and (5) 27 features that have the same tendency and which have been answered by the participants' supervisors, because they are mainly not under the most informative feature subsets (although *total clicks* is significant in the statistical comparison).

We report the two best F1-scores and accuracy scores for each data set as well as the baseline, as can be seen in Table 4.12. We outperform our basic baseline (DummyClassifier) for all data sets. The best F1-score, *0.75*, is achieved for both languages, the DE and ES data sets. DE uses 5 features with RF and ES uses ETC with 20 features. The second best F1-score, *0.74*, is achieved with the DE data set using 5 features and RFW. The best accuracy, *0.74*, is achieved with RF while the second best of *0.73* is achieved with RFW, both in the DE data set using just 5 features.

For ES, the best F1-score is also *0.75* with ETC and the selection of 20 features. The second best F1-score for ES is *0.72* with RF and a selection of 5 features. The F1-score is reduced by 0.10 when combining the two data sets (DE and ES), since the best F1-score for ALL is *0.65* using GB and 20 features. The second best F1-score for ALL is *0.63* with GB and 5 features. For ES, the best

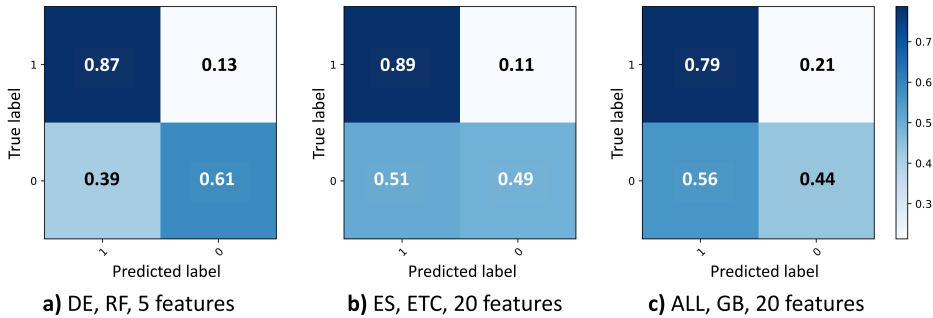


Figure 4.6: Normalised confusion matrix from the three best results (F1-score and accuracy): **a)** *DE*, 5 features with *RF*; **b)** *ES*, 20 features with *ETC*; and **c)** *ALL*, 20 features with *GB*.

accuracy is 0.69 with *ETC* and the selection of 20 features. The second best accuracy for *ES* is 0.65 with *RF* and a selection of 5 features. The accuracy is reduced by nearly 0.10 when combining the two data sets (*DE* and *ES*), since the best accuracy for *ALL* is 0.61 using *GB* and 20 features. The second best accuracy for *ALL* is 0.59 with *GB* and 5 features. This shows that there are differences across languages.

The normalised confusion matrix (see Figure 4.6) does not show over-fitting for the best obtain results for *DE*, *ES* and *ALL*.

The reduction of features improves the accuracy for *DE* but not consistently for *ES* and *ALL*, as can be seen for the different classifiers and data sets in Figure 4.7. For example, reducing the features for *DE* improves the accuracy for *ET*, *RF*, and *RFW*, but not for *GB*. For *ES*, the accuracy improves only for *RF* and stagnates for *RFW* when reducing the number of features, otherwise the accuracy inverts for *ETC* and *GB*. For the data set *ALL*, *RFW* and *RF* improve but *ETC* and *GB* decrease.

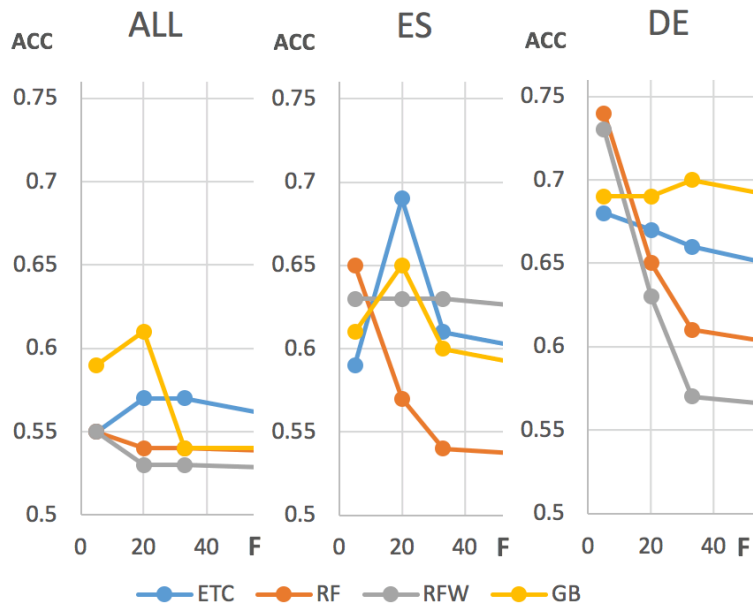


Figure 4.7: The plot shows the relation of accuracy to features for all classifiers in the data set ALL (left), ES (middle) and DE (right).

## 4.9 Discussion

Most children with dyslexia show a varying severity of deficits in more than one area [14], which makes dyslexia more a spectrum than a binary disorder. Additionally, we rely on current diagnostic tools (*e.g.*, DRT [51, 148]) to select our participant groups, which do not yet represent the diversity of people with dyslexia. We accept that our participants have a high variance because of the measurement of our current diagnostic tools and the spectrum dyslexia has.

In the following sections we will discuss the results from the statistical analysis 4.7 and the machine learning prediction 4.8.

### 4.9.1 Group Comparison

The measurement data taken from the game *MusVis* show that Spanish participants with dyslexia behave differently than their control group (R1). Differences can be reported for the auditory game part for: *4th click interval*, *6th click interval*, *duration*, and *average click time*. For the visual part, the following measurements can be reported as indicators: *total clicks*, *time to the first click*, *hits*, and *efficiency*. Besides, similar tendencies can be reported for the variables of the visual part: *total clicks and misses* (see Table 4.10).

We can show with our results over all languages that the effect for each measurement is confirmed even if we cannot draw strong conclusions about our sample size on the comparison of German vs. Spanish speaking participants. Spanish has eight significant indicators and we expected to reproduce the same number of significant indicators with more German participants.

In general, all participants found the game easy to understand, and only children at the age of 12 complained about missing challenges. The amount of positive feedback and engagement of all age groups let us conclude that the game mechanics and components applied are also positive for perceiving *MusVis* as a game and not as a test.

Dyslexia is known to be present across different languages and cultures [2]. The assumption that the tendencies for the indicators are similar over all languages cannot be proven for all indicators in our study (e.g., German participants with dyslexia start to click faster than the Spanish participants compared to their language control group in the auditory part). We can exclude external factors such as different applications or study setups as possible influences on this opposite tendency. According to the results, we may have to assume that not all indicators for dyslexia are language-independent and that some have cultural dependencies, or we have *omitted variable bias*. To confirm this assumption, we will need to obtain larger numbers of participants



for both language groups (Spanish and German) or investigate further measurements (indicators).

The variables *time to first click (visual and auditory)* and *total number of clicks (visual and auditory)* provide dependencies of the game content and game design. Otherwise, we could not explain the trend difference between the auditory and visual parts for *total number of clicks* (i.e., *total clicks* for visual is significantly different than for auditory). Additionally, the analysis of the auditory game part presents two limitations: (1) participants could select a correct pair by chance, and (2) participants could click through the game board without listening to the sounds.

Children with dyslexia are detected by their slower reading or spelling **error rate** [23, 140] as further explained in the Chapter 2. Therefore, we designed our game with content that is known to be difficult to differentiate for children with dyslexia to measure the errors and duration. Nevertheless, from previous literature we knew that children with dyslexia do not make more mistakes in games than the control group [127]. We can confirm that *misses* did not reveal significant differences for German or Spanish either. It might be possible that we cannot compare errors in reading and writing with errors in this type of game. Then, we cannot explain (yet) why the Spanish control group made more mistakes than the Spanish group with dyslexia. It might also be possible that participants with dyslexia show generally different behavior that is separated from the content but depends on the gameplay. Spanish children without dyslexia take significantly more time to find all pairs and finish the auditory game part. Children without dyslexia take more time before they *click the first time* (visual) for all languages. This might be due to the time they need to **process the given auditory information** [149] or recall the auditory and visual information from short-term memory [47]. However, participants with dyslexia from the German group are nearly as fast as the control group in finding all pairs (auditory) which might be due to **cultural differences**.

The auditory and visual cues are designed on purpose to be more difficult to process for people with dyslexia than without. Therefore, children with dyslexia are expected to need more time (duration), which might be due to a **less distinctive encoding of prosody** [47] and is in line with the indicator of slower reading. Considering that children with dyslexia need more time to process information, we observe this behavior as well for our indicators. For example, participants with dyslexia from the Spanish group take more time on the *4th click interval* and also on the *average click time* compared to the control group. Both results are significant and have large effect sizes of 0.7 and 0.8, so we can estimate what the effects would be in the whole population [37].

A person with dyslexia has difficulties with reading and writing independent of the mother tongue, which also appear when learning a second language [57, 90]. The analysis of errors from children with dyslexia show similar error categories for Spanish, English [126], and German [118], revealing similarities of perception between the languages.

Our results suggest that we can measure a significant difference on four indicators for the visual game with the same tendency between Spanish, German, English, and Catalan (R1). This means that people with dyslexia might perceive our visual game content similarly, independent of the mother tongue. Further research needs to be done to confirm the results, but this validation study provides strong evidence that it will be possible to screen dyslexia with our content, approach, and game design using the same language-independent content for different languages.

#### 4.9.2 Screening Differences

Our approach aims to screen dyslexia with indicators that do not require linguistic knowledge. These indicators are probably not as strong or visible as the reading and spelling mistakes of children with dyslexia. Therefore, we consider our results (highest accu-

racy of 0.74 and highest F1-scores of 0.75) for German with Random Forest as a promising way to predict dyslexia using language-independent auditory and visual content (R2). Having an early indication of dyslexia before spelling or reading errors appear can have a positive impact on the child's development, as we can intervene earlier in her/his education.

Therefore, we aim to optimize the Recall and F1-score by finding as many participants with dyslexia as possible.

We have set ourselves this goal because early detection in a person with dyslexia has a greater positive effect on the person with dyslexia than a misjudgement in a person without dyslexia. If a person with dyslexia is not discovered (early), they are prone to face additional issues such as anxiety, sadness and decreased attention [138]. Also, a person with dyslexia needs around two years to compensate for their reading and spelling difficulties. Early treatment among children at risk of dyslexia as well as children without dyslexia can serve as both a preventive measure and an early stimulation of literacy skills.

Our results support the hypothesis that dyslexia cannot be reduced to one cause, but is rather a combination of characteristics [27]. The equal distribution of auditory and visual features in the informative features ranking supports the hypothesis of dyslexia being related to auditory and visual perception. We might be able to measure stronger effects when we design visual and auditory cues that have more attributes related to dyslexia.

The ALL data set reached *only* an accuracy of 0.61, which might be due to the following reasons (R2). First, the informative features for each data set are different from each other, which indicates different informativeness in German and Spanish. Combining the data sets into ALL probably adds noise for the prediction, which results in a lower accuracy. The noise might be that features are not as informative anymore because they cancel each other out as they are highly correlated.

In addition, reducing the features only to the features with the same tendency as used for the statistical analysis did not reveal any improvement, which supports the hypothesis that features in ALL cancel each other out.

The results of our current game measures with 313 participants confirm differences in the behavior of Spanish vs. German participants (*i.e.*, (1) seven significant dependent variables in Spanish vs. none in German and (2) only two dependent variables with the same tendency over all languages).

These results might be explained with the concept of bilingualism. It is argued that a person who speaks more than one language has more knowledge of their first language than a monolingual person [73], and it is unclear whether this also has an influence on “how people perceive differences as well”. Additionally, dyslexia detection differences are reported for transparent (like Spanish) vs. deep (like English) orthographies (quoted after [125]). In a transparent orthography mainly a single grapheme (letter) corresponds to a single phoneme (sound) and dyslexia is reported to be more distinct in deep orthographies.

If so, this might explain the difference we have in the significance for the statistical analysis as well as the tendency of values, and the need for separate models to predict dyslexia for our German vs. Spanish data set (Spanish has bilingual participants).

Overall, having fewer features improves the accuracy, but this is less so when we run experiments for ALL or ES. There, the influence of the different informative features for ES and DE seem to cancel each other out. The high correlation between features would explain why, for example, ALL taking into account 27 features (GB) performs no better than using 20 features (GB) from the ALL data set. The fact that the accuracy does not increase when more features are used supports the argument that features are highly correlated.

As described before, small data can help for understanding the data and results better. In our case, we see that ALL does not per-

form as well as ES or DE. This is probably due to the facts described above (e.g., bilingualism, features canceling each other, English-speaking participants). The prediction for dyslexia is therefore possible with the data taken from the same game, but needs different models for the prediction in different languages as was proposed by [6], something that made sense in retrospect (R2).

# Screening Dyslexia adding Generic Content

---

## 5.1 Introduction

In Chapter 4, we showed how a game can be used to screen for dyslexia in English, German and Spanish with auditory and visual cues. We addressed language-independent screening with mainly a desktop game in order to later integrate younger children who do not yet have any reading or writing skills. Based on these results, we redesigned the gameplay and added cues with generic visual and auditory content related to various indicators. Hence, here we present a game to study the relation of language-related vs. generic game content to screen children with dyslexia.

Our main contributions are the added language-independent generic content related to various indicators, the relaunched auditory gameplay, and the prediction with interaction data from a web-based game using machine learning. The contributions of this chapter can be summarized as follows:

- The game design presents how the new auditory game part, the new language-independent auditory content as well as the added generic content for auditory and visual related to various indicators is designed (see Section 5.3).
- The experimental design setup, predictive model setup and the results for the evaluation of the web-based game provide insights into how participants with and without dyslexia differ from each other and how language-related and generic content with various indicators can be used for screening dyslexia (see Sections 5.4, 5.5, and 5.6).
- The discussion shows how children with and without dyslexia can be distinguished (R1), for example, with nine significant variables for German (see Section 5.7). Additionally, we discuss how the new generic visual and auditory content related to various indicators perform best (R3). We reached the best balanced accuracy and F1-score for all languages (0.77,0.75) and for German (0.67,0.74) with generic auditory cues using Extra Trees Classifier (see Section 5.7).

Part of the content of Section 5.3 was published in [113].

## 5.2 Methodology

Generally, we follow the methodology explained in Chapter 3 and point out here the important and case-specific steps for this application and study.

Consequently, we used the study and observational results as well as participant feedback and game data of *MusVis* to develop a new application, *DGames*, with the following requirements: (a) add generic content; (b) provide auditory content that is appropriately perceivable by pre-readers; (c) simplify the auditory gameplay; and (d) use input methods that are adequate for pre-readers. Hence,

we changed the gameplay and content of the application as well as dependent variables for the experimental design. We aimed for easier gameplay and an auditory gameplay comparable with the visual game part. The content needs to incorporate cues related to various deficits known to be connected to dyslexia as well as generic content and content more perceptible for pre-readers. The major changes made to achieve the requirements are described in Section 5.3.

As the comorbidity is a challenge for dyslexia research (as explained in Chapter 2 and 3), we added further questions regarding possible related diagnoses of the participants (e.g., ADHD diagnosis, hearing limitations). We used internal feedback loops from human-computer interaction researchers to ensure the quality of the game as we already evaluated the gameplay of *MusVis*.

Due to limited resources (e.g., time, personnel, language skills), we could only collect a sufficient number of participants for German (see Section 5.4). We combined all participants in an additional data set to decide if collecting further participants in future studies might have positive prediction results for the other languages (English, Spanish).

We evaluated our indicators with the statistical analysis and compared the statistical results between *MusVis* and *DGames*. The statistical and prediction results are presented in Section 5.6 and the discussion in Section 5.7.

## 5.3 Game Design

The game *DGames* is similar to the game *MusVis* [115] and a demo is available at <http://bit.ly/DGamesEN>. Only the interaction for the visual part and eight visual cues are duplicated from *MusVis*. We designed the game *DGames* with the experience from the previous study and implementation of *MusVis* [115, 117]. All changes are reported below.



Although children aged 7 to 12 and their parents gave very positive feedback on the gameplay and content, parents of pre-readers described troubles. Parents reported that their pre-readers had difficulties in understanding the gameplay and distinguishing the very short and similar sounds of the auditory part. Most of them quit the game because of that. An example quote from a dad of a boy (4 years) shows this: "He was overwhelmed by the game. He could not distinguish the sounds and just touched randomly on any card." Also, the input method (computer mouse) was not adequate for younger children.

To be able to target pre-readers, we changed the focus in the implementation from a desktop computer to a tablet device. We also completely recreated the auditory content and interaction. Additionally, results from *MusVis* and *Dyctective* [127, 129] showed that children with dyslexia (CWD) did not make more mistakes while playing games, in spite of the fact that CWD are diagnosed by the number of written errors they make. Therefore, we design more generic auditory and visual content to be less linguistic (called *generic*) to evaluate the influence on the prediction.

The auditory game cues from *MusVis* were related to one acoustic parameter. Since the variables for the auditory part of *MusVis* are mainly not significant, we assume that cues related to various deficits (as done for visual) will increase the chances of measuring a difference between groups. Since the short-term memory capacity shows a strong correlation with the search variable [10], our games are both designed as search games. Each game part (auditory and visual) has 16 rounds which are counter-balanced with *Latin Squares* [37].

We aim to address users' motivation for both game parts with the following game mechanics, which are similar to the ones used for *MusVis*: rewards (points), feedback (instant feedback), or challenges (time limit). We identified these game mechanics as techniques to motivate through emotional engagement and visualize users' progress (as explained in Chapter 3).

### 5.3.1 Auditory Cues adding Generic Content

The new auditory cues are designed with the knowledge of attributes used in previous literature and the new analysis of the German errors resource (see Table 5.1). Our analysis of German error words [109, 118] compared with the existing analysis of Spanish and English [126] found similarities between the languages as described in Appendix A.2. The analysis of error words on different languages needs future work to make a comprehensive assessment, but this is not the main scope of this thesis.

We mapped the attributes explained in Table 5.1 to the stages in Table 5.2. Additionally, Table 5.2 shows the relations between our designed auditory types and the literature that provides evidence for distinguishing a person with dyslexia.

The auditory part has for each stage/round a new auditory type: **substitution**, **omission**, **structure**, **phoneme** (one Spanish and one German vowel; Spanish consonant), **confusion** (twice Spanish and German; four times English), **combinations**, and **rhythm**. Each auditory type has one auditory cue target and three auditory cue distractors.

Auditory types are related to linguistic features when using the pronunciation of letters or the confusion of words. For example, we created cues from the vowel pronunciation of letters using the Mac OS High Sierra 10.13.6 voice for the different languages, e.g., Spanish and German. For example, we know from the analysis of mistakes from children with dyslexia in Spanish that vowel errors are the most frequent in the substitution category [126].

The auditory game round has two phases: (1) remembering the target audio cue; and (2) finding the target audio cue among a collection of audio cues. In the first phase, the children click on the *play* button and can listen to the auditory cue target as often as they like (see Figure 5.1, a). In the second phase, a row of four buttons is displayed (see Figure 5.1, b) and automatically the assigned auditory cues for each button are played one after another.

<b>Key</b>	<b>Name</b>	<b>Description</b>
<b>B</b>	Beginning	70% of the spelling errors are at the third position of a word for German and Spanish [109, 126].
<b>L</b>	Length	The average word length for German and Spanish is just above 7 letters [109, 126].
<b>Si</b>	Simple	For 73.3% of the analyzed words for Spanish the Damerau-Levenshtein distance was one, which means that only single mistakes were made [126]. For German it is even 81.3% [109].
<b>Su</b>	Substitution	The error category <i>Substitution</i> (exchanging a letter for another one) is frequent for German, English and Spanish [109, 126].
<b>O</b>	Omission	The error category <i>Omission</i> (leaving a letter out) is frequent in German [109].
<b>St</b>	Structure	CWD find it more difficult to recall a target item with a similar prosodic structure [47].
<b>Pst</b>	Phonological short-term memory	CWD showed difficulties in the phonological short-term memory [47].
<b>Sip</b>	Short-interval perception	Copying and discrimination tasks are used to predict phonological awareness [85].
<b>Pm</b>	Pitch modulation	CWD have difficulties in processing pitch patterns [133].
<b>Cb</b>	Combinations	Discrimination of rise time is related to language processing [48].
<b>C</b>	Complexity	CWD have difficulties with <i>the phonological similarity effect</i> and the <i>phonological neighbourhood</i> when long memory spans are used [47]. English has a greater percentage of multi-errors compared to Spanish [126].

Table 5.1: Description of the auditory attributes for *DGames*.

Content	Related		Generic				
	Phon.	Conf.	Comb.	Omi.	Rhyt.	Struc.	Subs.
Type							
Total Stages	3	8	1	1	1	1	1
<b>Attri.</b>							
B	✓		✓	✓		✓	✓
L	✓	✓	✓	✓		✓	✓
Si	✓			✓		✓	✓
Su	✓		✓				✓
O	✓		✓	✓	✓		
St	✓	✓	✓	✓	✓	✓	✓
Pst	✓	✓	✓	✓	✓	✓	✓
Sip		✓	✓		✓		
Pm	✓	✓	✓			✓	✓
Cb			✓		✓		
C	✓		✓		✓		

Table 5.2: Mapping of the evidence from literature to distinguish a person of dyslexia to design the auditory type for each stage of *DGames*.

The buttons are disabled until the auto-play is done to ensure the children listen to all auditory cues. In order to distract the player, the first button/auditory cue is never the auditory cue target. The order of auditory cues is randomly assigned and starts always from left to right. With the *Play all sounds again* button, the children can listen to all cues as often as they like.

### 5.3.2 Visual Cues adding Generic Content

The main changes from *MusVis* [115] are (a) adding non-linguistic content (called *generic*), (b) tablet adaptation (e.g., double click), and (c) video introduction to the game.

The visual part has 8 stages and 16 rounds. Each stage has two rounds with first a 4-squared (see Figure 5.2, b) and then a

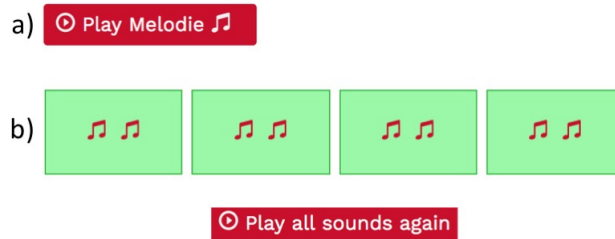


Figure 5.1: Example of the auditory part of the game *DGames* with the priming of the target cue (a) and then the distractors for each auditory cue (b).

9-squared design (see Figure 5.2, c). Each stage is assigned to one visual type (**symbol**, **z**, **rectangle**, **face**, **fruit**, **kitchen**, **plant**, and **animal**). Four visual cues for each stage are presented (see Figure 5.3, where the first four visual types are duplicated from *MusVis* [115]).

The visual game round has two phases: (1) remembering the target visual cue; and (2) finding the target visual cue among a collection of visual cues. In the first phase, the participant has 3 seconds to memorize the assigned target visual cue for the stage (see an example in Figure 5.2, a). This is the visual target cue the participant has to find in the second phase of the round.

In the second phase of the round, the target and all three distractors are displayed in either the 4-squared design or 9-squared design for 15 seconds. The participant needs to find and click on the target visual cue before the time is up. Within the 15 seconds, the squared design is updated every time the participant clicks on a visual cue. The visual cues are randomly assigned within the squared design. The target and all three distractors are displayed in the 4-squared design. In the 9-squared design, the target is displayed twice along with distractors two and three. Only distractor one is displayed three times.

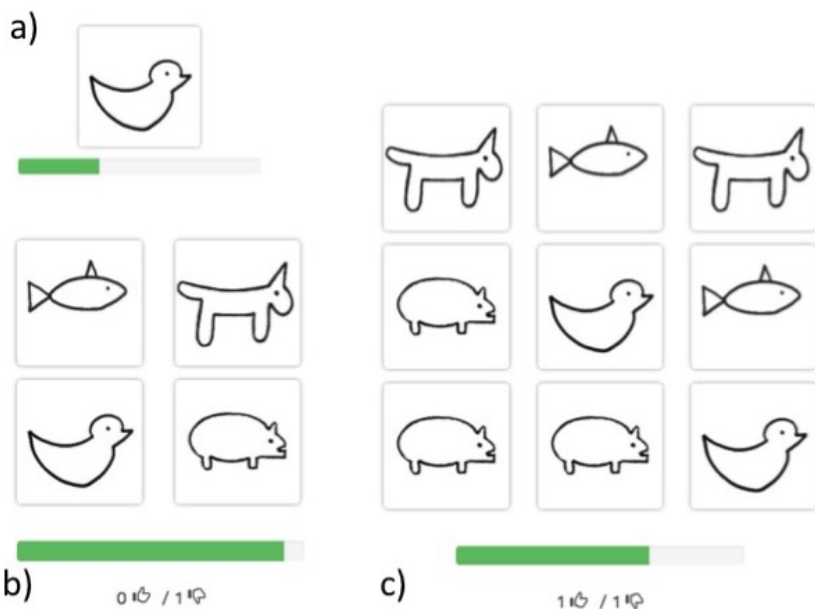


Figure 5.2: Example of the visual part of the game *DGames* with the priming of the target cue *animal* (a) and then the four-squared (b) and (c) nine-squared design including the distractors for each *animal*.

### 5.3.3 User Interface

As for *MusVis*, we avoid distractions or influences on the users' behavior with a consistent user interface (e.g, font size, color or shape) and support the readability of text for participants' supervisors with a large font size (minimum 18 points) [124].

The interactive elements (cards to be clicked within the game) are large enough to be touched easily. The interactive elements (sound cards/squares) are presented the same way within each game and do not differ in color or shape to avoid differences in perception.

































	Related				Generic			
	symbol	z	rectangle	face	fruit	kitchen	plant	animal
target								
distractor 1								
distractor 2								
distractor 3								

Figure 5.3: Overview of the designed related-linguistic (called *related*) and non-linguistic (called *generic*) cues. The figure shows the target cue (top) and distractor cues (below) for the eight different stages (*symbol, z, rectangle, face, fruit, kitchen, plant, animal*) of the visual part of the game *DGames*.

### 5.3.4 Implementation

Both game parts are developed as a web application using JavaScript, jQuery, CSS, HTML5 and a back-end with a PHP server and a MySQL database to make the game easily adaptable for different devices. The visual part is also implemented with Angular.

Because of the web implementation technique, we were reminded that a double click on a web application generally *zooms* the application on a tablet. As young children were observed touching the application very quickly and triggering the *zoom*-effect, this caused interruptions while playing and was not coherent with the experience of using a native tablet application. Therefore, we used a *viewport meta tag* to control the layout settings for mobile devices.

All instructions within the game are presented with video or audio media to address pre-readers. *Android* prevents by default *automatic play of sound or video* and asks for user interaction. Therefore, we designed the whole game with a sequence that starts with user interaction followed by audio sounds.

## 5.4 Experimental Design Setup

We designed our application with the human-centered design framework [67] to gather data for the prediction of dyslexia with an experimental study design [37]. In our *within-subject design* [37], all participants played all 16 rounds of each game part.

As for *MusVis*, we also collect data from children ages 7 to 12. First, this is because at this age children should already be diagnosed with dyslexia. Even if a treatment is already being performed, children should still show signs of dyslexia. Second, we do this because it allows us to compare the results with *MusVis* where we collected data from the same age range. As described in Section 3.6.2 the study has been approved.

### 5.4.1 Procedure

We followed the same approach as described for *MusVis* in Section 4.5.1. Apart from the following: DGames has a video instruction also for the visual game. Participants watched the short video instructions for the visual game part and played the visual game. Following, the participants watched the short video instructions for the auditory game part and played the auditory game.

### 5.4.2 Participants

We have the following exclusion and inclusion criteria: We *excluded* participants if they did *not* report any of the following at-



Data set	N	Dyslexia				Control			
		N	$\overline{age}$	female	male	N	$\overline{age}$	female	male
DE	120	36	9.1	17	19	84	10	46	38
ALL	137	51	9.7	23	28	86	8.8	48	38

Table 5.3: Overview of the participants per data set.

tributes: gender (3.5%), age (7.1%), or diagnosis status (8.5%). Only participants who played all 16 rounds of each game are included in the analysis. Participants with dyslexia were mainly recruited over (support groups on) social media calls, learning centers and our own participant pool. The control groups were recruited mostly with the collaboration of schools from the northern part of Germany (Schleswig-Holstein and Lower Saxony State).

The participant call raised attention for parents who either did not know whether their child had dyslexia (18.3%) or suspected their child had dyslexia but did not have an official diagnosis (9.8%), *e.g.*, from a medical doctor. To have a precise small data set and a binary classification to simplify the prediction, we only considered participants with an official diagnosis (dyslexia group) or no sign of dyslexia (control group).

We separated our data in two data sets (see Table 5.3): One data set with German participants (DE,  $n = 120$ ) and the other with all languages. The DE data set is constrained to be as precise as possible (*e.g.*, one language, no missing values, and restricted age range). This means we took out optional questions when only a few participants answered. We cannot completely balance our participant groups for dyslexia ( $n = 36$ ) and control ( $n = 84$ ) in DE due to limited resources, which we will address in our analysis.

The other data set includes all languages (ALL,  $n = 137$ ) where we also added participants that used English ( $n = 2$ , 1 dyslexia and 1 control) and Spanish ( $n = 15$ , 14 dyslexia, 1 control) game instructions. We use the ALL data with the same constraints as DE to explore the influence on the prediction, apart from balanced groups of participants for Spanish (mainly participants with dyslexia) and

Data set	N	Biling.Δ in %	Dyslexia				Control					
			N	One Lang.* N	in %	Biling.Δ N	in %	N	One Lang.* N	in %	Biling.Δ N	in %
DE	120	19.16	36	31	86.11	5	13.89	84	66	78.57	18	21.43
ALL	137	22.63	51	40	78.43	11	21.57	86	66	76.74	20	23.26

Table 5.4: Overview of bilingualism per data set. \*One Language; Δ Bilingualism.

different languages. Participants ranged in age from 7 to 12 years.

Our ALL data set ( $n = 137$ ,  $\overline{age} = 9.2$ ,  $sd = 1.5$ ) contains 51 participants with dyslexia (23 female, 28 male,  $\overline{age} = 9.7$ ) and 86 as control (48 female, 38 male,  $\overline{age} = 8.8$ ).

Our German data set ( $n = 120$ ,  $\overline{age} = 9.1$ ,  $sd = 1.5$ ) includes 36 participants with dyslexia (17 female, 19 male,  $\overline{age} = 10$ ) and 84 as control (46 female, 38 male,  $\overline{age} = 8.8$ ).

Participants played the game in English, German, or Spanish depending on their native language. We provide an overview in Table 5.4 on the answers for one language (One Lang.) and reported bilingualism (Biling.) from the demographic questionnaire. Our data sets contain bilingual participants (ALL  $\sim 22.63\%$ , DE  $\sim 19.16\%$ , *Biling.*).

The participants with dyslexia mainly reported one native language for ALL ( $\sim 78.43\%$ ,  $n = 40$ ) and DE ( $\sim 86.11\%$ ,  $n = 31$ ). For ALL,  $\sim 21.57\%$  of the participants with dyslexia reported bilingualism ( $n = 9$  a second and  $n = 2$  a third language). Only  $\sim 13.89\%$  of the DE participants with dyslexia reported bilingualism ( $n = 4$  a second and  $n = 1$  a third language). The participants without dyslexia reported one native language for ALL with  $\sim 76.74\%$  ( $n = 66$ ) and for DE with  $\sim 78.57\%$  ( $n = 66$ ). For ALL control,  $\sim 23.26\%$  of the participants reported bilingualism ( $n = 13$  a second,  $n = 5$  a third, and  $n = 2$  a fourth language). For DE control,  $\sim 21.43\%$  of the participants reported bilingualism ( $n = 11$  a second,  $n = 5$  a third, and  $n = 2$  a fourth language). Bilingual participants used the language they reported to be more comfortable with for the instructions of the game.

<b>Part</b>	<b>Dependent variables</b>
<b>Auditory</b>	Number of repeats, correct answers, wrong answers, duration on per round, thinking time of the participants before responding.
<b>Visual</b>	Total clicks, correct answers, wrong answers, 1st click interval, 2nd click interval, 3rd click interval, 4th click interval, 5th click interval, 6th click interval, time last click, accuracy, efficiency and effect.

Table 5.5: Overview of the dependent variables used for the statistical comparison.

### 5.4.3 Dependent Variables and Features

We conducted an online user study to collect participants' responses with the demographic questionnaire and the dependent variables while playing the web game. We use the dependent variables from children with and without dyslexia for the statistical comparison and also as input (features) for the machine learning classifiers. First, the statistical analysis is performed using the dependent variables presented in Table 5.5.

Next, we selected our features for our predictive models from the collected data taking into account the following sources: questionnaire, technical meta data, and game measures.

The features 1 to 10 were answered with the online questions by the subject's supervisor (see Table 5.6). The technical meta-data, Features 11 to 13, was collected with the PHP variable from the browser while the user was participating in the user study (see Table 5.6).

We would like to further elaborate on our subject features *language*, *class level* and *device*. We select only one subset (DE) with the feature *language*. Our game instructions are available in three languages. But due to limited resources (time and personnel), we collected only sufficient amount of participants for the German data set (English,  $n = 2$  and Spanish,  $n = 15$ ). The feature *class level*

ranges from 0 to 8 and describes in which year of education the participant is. The integer value corresponds to the main model of primary and lower secondary education in Europe [32]. As the influence of the input method (*computer* vs. *tablet*) was in *MusVis* (Chapter 4) positive for the prediction, we included for the analysis both input methods and collected further technical metadata such as *operating system* and *browser*.

We excluded a few questions because the participants' answers could not be solidly summarized for the feature selection. For example, the questionnaire responses to school grades for language or mathematics subjects were different and could not be attributed to a uniform result, such as 2, ++, -, *remarkable*.

The data set has 429 features per subject, where the 160 variables from the auditory part are features 14 to 173 (Table 5.7) and the 256 variables for the visual part are 174 to 429 (in Tables 5.8 and 5.9). A game variable corresponds to either the auditory or visual round.

We would like to further elaborate on our auditory features *duration per round*, *thinking duration* and *button position*: The function *duration per round* starts with the beginning of the game round, which is the listening of the target queue, and ends when the participant has made a choice in the second phase (that is, a click on a button).

We elaborate on the gameplay to explain the feature *thinking duration*. Buttons are assigned to auditory cues. After the second phase starts, auditory cues are automatically played one after another. The buttons are disabled until the auto-play is done to ensure children listen to all auditory cues.

*Thinking duration* lasts from the time the buttons are no longer disabled until the participant chooses (clicks on) a button.

The feature *button position* represents the position of the button within the row of the four buttons, starting from the left.

We would like to further elaborate on our visual features for the different click intervals. From *MusVis* (Chapter 4), we know that

<b>Participant features</b>	<b>Feature description</b>
<b>1</b> Age	It ranges from 7 to 12 years old.
<b>2</b> Gender	It is a binary feature either with <i>female</i> or <i>male</i> value.
<b>3</b> Language	It indicates the instruction language, <i>Spanish, German or English</i> .
<b>4</b> Native Language	It indicates if the language used for the instructions is the first language of the participant, being <i>Yes</i> or <i>No</i> .
<b>5</b> Number of languages	It describes the number of languages a participant reported knowing, ranging from 1 to 4 languages.
<b>6</b> Class level	It describes the class level of the participant, being an integer ranging from 1 to 8.
<b>7</b> Hearing Limitations	It indicates the hearing limitation the participant reported, being <i>No limitations, little limitations, or limitations</i> .
<b>8</b> Fun	It indicates the expressed fun mentioned in the feedback question, being <i>No fun, little fun, or fun</i> .
<b>9</b> Difficulty Level	It indicates the expressed level of difficulty for the game mentioned in the feedback question, being <i>Not challenging, middle challenging, challenging</i> .
<b>10</b> Instrument	It indicates if a participant plays a musical instrument, being <i>No, Yes, less than 6 months</i> or <i>Yes, over 6 months</i> .
<b>11</b> Device	It is the device the participant used, which is a binary feature with the <i>Computer</i> or <i>Tablet</i> value.
<b>12</b> Operating system	It describes the operating system the participant used, being <i>Mac OS, Windows, Android</i> or <i>Linux</i> .
<b>13</b> Browser	It is the browser the participant used, being <i>Safari, Chrome, Edge, Firefox, Opera</i> or <i>Internet Explorer</i> .

Table 5.6: Description of the participant features.

Auditory features	Feature description
<b>14–29</b> Number of repeats	It counts the number of times a participant listened to the instructions, being an integer.
<b>30–45</b> Duration per round.	It is the duration of a round, being <i>milliseconds</i> .
<b>46–61</b> Thinking duration	It is the duration after the buttons are not disabled anymore until the participant chose (clicked) an auditory cue in the second phase of the game round, being <i>milliseconds</i> .
<b>62–77</b> Number of repeats target melody	It is counting the number of times a participant listened to the target auditory cue, being a <i>integer</i> .
<b>78–93</b> Correct answers.	It is a binary feature either with <i>wrong</i> or <i>correct</i> value.
<b>94–109</b> Wrong answers.	It is a binary feature either with <i>wrong</i> or <i>correct</i> value.
<b>110–125</b> Total correct answers.	It indicates the number of correct answers for the previous rounds by summing the correct answers from all previous auditory rounds, being a <i>integer</i> .
<b>126–141</b> Total wrong answers.	It indicates the number of wrong answers for the previous rounds by summing the wrong answers from all previous auditory rounds, being a <i>integer</i> .
<b>142–157</b> Cue	It indicates the participant's click choice, being <i>target</i> , <i>distractor 1</i> , <i>distractor 2</i> , or <i>distractor 3</i> .
<b>158–173</b> Button position	It describes the position of the clicked button, being <i>left</i> , <i>middle-left</i> , <i>middle-right</i> , or <i>right</i> .

Table 5.7: Description of the auditory features.

Visual features	Feature description
<b>174–189</b> 1st click interval.	It is the duration between the start of the second phase and the first click of the participant on a visual cue, being <i>milliseconds</i> .
<b>190–205</b> 2nd click interval.	It is the duration between the first and second clicks of the participant on a visual cue, being <i>milliseconds</i> .
<b>206–221</b> 3rd click interval.	It is the duration between the second and third clicks of the participant on a visual cue, being <i>milliseconds</i> .
<b>222–237</b> 4th click interval.	It is the duration between the third and fourth clicks of the participant on a visual cue, being <i>milliseconds</i> .
<b>238–253</b> 5th click interval.	It is the duration between the fourth and fifth clicks of the participant on a visual cue, being <i>milliseconds</i> .
<b>254–269</b> 6th click interval.	It is the duration between the fifth and sixth clicks of the participant on a visual cue, being <i>milliseconds</i> .
<b>270–285</b> Time last click	It is the time of the last click within a game round in the second phase, being <i>milliseconds</i> .
<b>286–301</b> Total clicks.	It is the number of total clicks within a game round, being an <i>integer</i> .
<b>302–317</b> Correct answers.	It is the number of hits or correct answers within a game round, being a <i>integer</i> .
<b>318–333</b> Wrong answers.	It is the number of wrong answers or non-correct answers within a game round, being a <i>integer</i> .

Table 5.8: Description of the visual features (part one).

Visual features	Feature description
<b>334–349</b> Distractor 1	It is the number of times distractor 1 is clicked within a round, being a <i>integer</i> .
<b>350–365</b> Distractor 2	It is the number of times distractor 2 is clicked within a round, being a <i>integer</i> .
<b>366–381</b> Distractor 3	It is the number of times distractor 3 is clicked within a round, being a <i>integer</i> .
<b>382–397</b> Efficiency	It is dividing the time of the last click by hits, being a <i>fractional value</i> .
<b>398–413</b> Accuracy	It is dividing the number of hits by the total number of clicks, being a <i>fractional value</i> .
<b>414–429</b> Effect	It is multiplying the number of hits by the total number of clicks, being a <i>integer</i> .

Table 5.9: Description of the visual features (part two).

participants with dyslexia have an average of 6.7 clicks per round [115]. We chose six intervals for DGames because we wanted to ensure meaningful values in the interval features and reduce the number of zero values due to missing clicks.

Next, we describe the model and feature selection.

## 5.5 Predictive Models Setup

In this section we present the machine learning for the ALL ( $n = 137$ ) and German ( $n = 120$ ) data sets. First we explain the choice of predictive models and then the choice of feature selection.

### 5.5.1 Model Selection

We used Random Forest (RF), Random Forest with class weights (RFW), Extra Trees (ETC), Gradient Boosting (GB), and the Dummy Classifier (Baseline), as in Section 4.6.1.



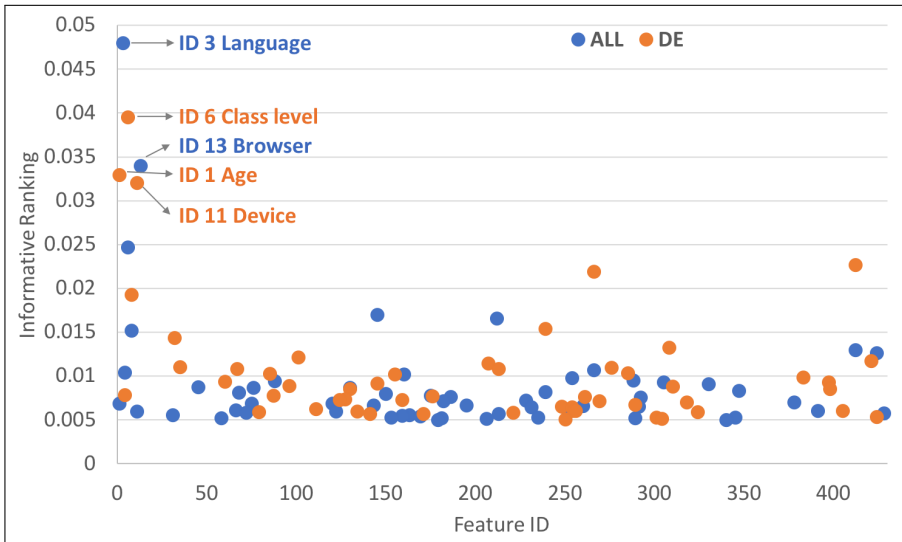


Figure 5.4: The plot shows the features ranking for the DE and ALL data sets with the highest-ranked features highlighted. Feature ID follows the index described in Section 5.4.3.

Following Section 3.7, we used a 10-fold cross validation and computed the *balanced accuracy* for our binary classification problem to deal with imbalanced data sets for ALL data set (dyslexia  $\sim$  37% vs. control  $\sim$  63%) and DE (dyslexia 30% vs. control 70%).

To compare different predictions we use the measures described in Section 3.7 and we used the feature selection as described in the next section.

### 5.5.2 Feature Selection

First, we ranked the most informative features with the *Extra Trees Classifier* (see Figure 5.4). The results show a step at the information score of 0.032: ALL 2 features and DE 3 features.

The two highest-ranked features of the ALL data set (*language and browser*) are different from the highest ranked features of the

Feature category	Number of features	Feature names
Auditory	3	cue 3, total number of wrong answers 4, button position 1.
Visual	7	accuracy 15, total clicks 4, effect 11, 3rd click interval 8, 5th click interval 2, 6th click interval 1, 6th click interval 13.
Subject	4	class level, age, fun, native language
Technical	1	device
<b>Total</b>	<b>15</b>	

Table 5.10: DE and ALL have 15 features in common among the highest-ranked informative features (DE  $n = 53$ , ALL  $n = 58$ ).

DE data set (*class level, age and device*). The comparison of the highest-ranked features (score over 0.005) reveals that the data sets have fewer features in common (15 features out of 58 for ALL and 53 for DE, see Table 5.10).

The game is designed with game content that should represent dyslexia, which means language-independent content. Since we have small data, the parameter optimization of predictive models can lead to over-fitting, so we use the feature selection to compare screening results. We explore the improvement of our measures for the predictive models with the subsets of features as described in Table 5.11. We address the danger of selecting the correct features [69] by taking into account knowledge of previous literature about the differences of children with and without dyslexia to avoid finding correlations by chance. All feature subsets include the subject features (1 to 13) except selected feature ID *informative*.

Selected features ID	Description
All features	All the 429 features.
Informative	They are the most informative features with a ranking score over 0.005 for ALL (58 features) and DE (53 features).
Auditory	Only the auditory features.
Auditory related	They are measures taken from the auditory game rounds where the game content was related to language.
Auditory generic	They are measures taken from the auditory game rounds where the game input was generic.
Visual	Only the visual features.
Visual related	They are measures taken from the visual game rounds where the game content was related to language.
Visual generic	They are measures taken from the visual game rounds where the game content was generic.

Table 5.11: Overview of the subsets of features used to compare the quality of the prediction.

## 5.6 Results

We present separately the results for the statistical analysis to find differences between children with and without dyslexia and for the predictive models to screen children’s risk of dyslexia.

### 5.6.1 Statistical Analysis

We compared the variables for our independent within-subject design only for German since we do not have enough participants for Spanish or English. Additionally, the groups (dyslexia/control) are not balanced, and we cannot investigate the tendency between par-

ticipants with dyslexia and the control group as we have done for *MusVis*. Therefore, we cannot select the dependent variables with the same tendency when combining data for the *all languages* data set. We report 9 significant dependent variables for German (DE): number of repeats, total clicks, correct answers, 1st click interval, 2nd click interval, 3rd click interval, time of last click, efficiency and effect. Since all variables were not normally distributed (*Shapiro-Wilk test*), we applied the *Mann-Whitney U* (also called independent Wilcoxon). We use the *Bonferroni correction* to avoid type I errors ( $p < 0.002$ ). We report dependent variables for DE in Table 5.12. The effect size is only reported for significant dependent variables and we use the classification defined by Cohen [37]. None of the significant dependent variables have a large effect ( $\geq 0.5$ ).

Next, we report the dependent variables for the auditory part.

**Number of repeats** has a small effect size and is significant ( $W = 327560.5, p = 6e - 9, d = 0.20$ ). Participants with dyslexia ( $m = 2.5, sd = 2.9$ ) listen on average more times to the target auditory cue than the control group ( $m = 2.1, sd = 2.2$ ).

**Correct answers** is not significant ( $W = 380064, p = 0.20$ ). Participants with dyslexia ( $m = 0.6, sd = 0.5$ ) have fewer correct answers than the control group ( $m = 0.7, sd = 0.5$ ).

**Wrong answers** is not significant ( $W = 379392, p = 0.20$ ). Participants with dyslexia ( $m = 0.4, sd = 0.5$ ) have more wrong answers than the control group ( $m = 0.3, sd = 0.5$ ).

**Duration** is not significant ( $W = 359961, p = 7e - 3$ ). Participants with dyslexia ( $m = 10.62s, sd = 28.70s$ ) take more time for a game round than the control group ( $m = 9.11s, sd = 6.94s$ ).

**Thinking time** is not significant ( $W = 358154.5, p = 5e - 3$ ). Participants with dyslexia ( $m = 1.16s, sd = 1.33s$ ) have a shorter thinking time than the control group ( $m = 1.31s, sd = 1.47s$ ).

Dependent variables	Control mean sd	Dyslexia mean sd	Mann-Whitney U W p-value	effect size
<b>Number of repeats</b>				
Correct answers	2.1 0.7	2.2 0.5	327561 38006	<b>6e-9</b> 0.20
Wrong answers	0.3 9.11s	0.5 6.94s	379392 359961	0.20 7e-3
Duration per round	1.31s	1.47s	358154	5e-3
Thinking time				
<b>Total clicks</b>	7.6	2.96	316124	<b>7e-11</b>
<b>Correct answers</b>	6.95	3.11	316955	<b>1e-10</b>
Wrong answers	0.66	1.23	382404	0.30
<b>1st click interval</b>	2.25s	1.39s	333016	<b>5e-7</b>
<b>2nd click interval</b>	2.07s	1.25s	344247	<b>6e-5</b>
<b>3rd click interval</b>	1.96s	1.07s	335961	<b>1e-6</b>
4th click interval	1.82s	1.05s	362120	0.01
5th click interval	1.60s	0.97s	370625	7e-2
6th click interval	1.31s	0.96s	384970	0.40
<b>Time last click</b>	13.79s	1.24s	354034	<b>2e-3</b>
Accuracy	0.90	0.19	380380	0.20
<b>Efficiency</b>	2.42	1.57	317231	<b>2e-10</b>
<b>Effect</b>	61.25	50.52	316518	<b>1e-10</b>

Table 5.12: Overview of all reported dependent variables for the auditory (top) and visual (below) part of the game *DGames* for DE ( $n = 120$ ). Significant results are in bold.

Next, we report the dependent variables for the visual part.

**Total clicks** has a medium effect size and is significant ( $W = 316124, p = 7e - 11, d = 0.31$ ). Participants with dyslexia ( $m = 8.55, sd = 3.18$ ) have more clicks within a game round than the control group ( $m = 7.6, sd = 2.96$ ).

**Correct answers** has a medium effect size and is significant ( $W = 316954.5, p = 1e - 10, d = 0.30$ ). Participants with dyslexia ( $m = 7.91, sd = 3.43$ ) have more correct answers than the control group ( $m = 6.95, sd = 3.11$ ).

**Wrong answers** is not significant ( $W = 382403.5, p = 0.30$ ). Participants with dyslexia ( $m = 0.64, sd = 1.26$ ) have fewer wrong answers than the control group ( $m = 0.66, sd = 1.23$ ).

**1st click interval** has a small effect size and is significant ( $W = 333016, p = 5e - 7, d = 0.18$ ). Participants with dyslexia ( $m = 2.01s, sd = 1.2s$ ) have a shorter interval than the control group ( $m = 2.25s, sd = 1.39s$ ).

**2nd click interval** has a small effect size and is significant ( $W = 344247, p = 6e - 5, d = 0.17$ ). Participants with dyslexia ( $m = 1.86s, sd = 1.11s$ ) have a shorter interval than the control group ( $m = 2.07s, sd = 1.25s$ ).

**3rd click interval** has a small effect size and is significant ( $W = 335960.5, p = 1e - 6, d = 0.20$ ). Participants with dyslexia ( $m = 1.76s, sd = 0.93s$ ) have a shorter interval than the control group ( $m = 1.96s, sd = 1.07s$ ).

**4th click interval** is not significant ( $W = 362120, p = 0.01$ ). Participants with dyslexia ( $m = 1.74s, sd = 1.2s$ ) have a shorter interval than the control group ( $m = 1.82s, sd = 1.05s$ ).

**5th click interval** is not significant ( $W = 370624.5, p = 7e - 2$ ). Participants with dyslexia ( $m = 1.56s, sd = 0.85s$ ) have a shorter interval than the control group ( $m = 1.60s, sd = 0.97s$ ).

**6th click interval** is not significant ( $W = 384970, p = 0.04$ ). Participants with dyslexia ( $m = 1.32s, sd = 0.85s$ ) have a longer interval than the control group ( $m = 1.31s, sd = 0.96s$ ).

**Time last click** has a small effect size and is significant ( $W = 354033.5, p = 2e - 3, d = 0.11$ ). For participants with dyslexia ( $m = 13.93s, sd = 1.23s$ ), the last click is later within a game round than for the control group ( $m = 13.79s, sd = 1.24s$ ).

**Accuracy** is not significant ( $W = 380380, p = 0.20$ ). Participants with dyslexia ( $m = 0.91, sd = 0.19$ ) are more accurate within a game round than the control group ( $m = 0.90, sd = 0.19$ ).

**Efficiency** has a small effect size and is significant ( $W = 317231, p = 2e - 10, d = 0.16$ ). Participants with dyslexia ( $m = 2.17, sd = 1.53$ ) are less efficient within a game round than the control group ( $m = 2.42, sd = 1.57$ ).

**Effect** has a medium effect size and is significant ( $W = 316518, p = 1e - 10, d = 0.31$ ). Participants with dyslexia ( $m = 77.81, sd = 61.90$ ) have a bigger effect within a game round than the control group ( $m = 61.25, sd = 50.52$ ).

From Table 5.12, we confirm nine significant dependent variables for *DGames* with German participants ( $n = 120$ ). Eight were dependent variables for the visual content (*total clicks, correct answers, 1st click interval, 2nd click interval, 3rd click interval, time last click, efficiency and effect*) and one was a dependent variable for auditory content (*number of repeats*).

## 5.6.2 Prediction using Machine Learning

We processed our data sets (ALL and DE) with different classifiers and different subsets of features as explained in Section 5.5.2.

The two best results for the F1-score and balanced accuracy obtained for each data set and feature selection as well as the baseline are presented in Table 5.13. Our results focus on the class dyslexia, e.g., F1-score, accuracy, recall, and precision.

We outperform our baseline (Dummy Classifier as described in Scikit-learn library version 0.21.2 [143]) for all data sets. For the ALL data set, we achieved the best two results for the F1-score

Sel. Feat.	ALL data set				DE data set					
	Clas.	Recall	Precis.	F1	Acc	Clas.	Recall	Precis.	F1	Acc.
All features	ETC	0.70	0.67	0.66	0.64	ETC	0.73	0.68	0.68	0.60
All features	GB	0.64	0.64	0.60	0.59	GB	0.70	0.65	0.66	0.59
All features	Baseline	0.63	0.39	0.48	0.50	Baseline	0.70	0.49	0.58	0.50
<b>Informative</b>	<b>ETC</b>	<b>0.77</b>	<b>0.81</b>	<b>0.75</b>	<b>0.73</b>	<b>ETC</b>	<b>0.79</b>	<b>0.78</b>	<b>0.74</b>	<b>0.65</b>
Informative	GB	0.75	0.79	0.73	0.70	GB	0.74	0.72	0.71	0.66
Informative	Baseline	0.63	0.39	0.48	0.50	Baseline	0.70	0.49	0.58	0.50
Auditory	RF	0.64	0.61	0.61	0.58	ETC	0.75	0.72	0.71	0.63
Auditory	RFW	0.65	0.64	0.61	0.58	RF	0.71	0.62	0.65	0.56
Auditory	Baseline	0.63	0.39	0.48	0.50	Baseline	0.70	0.49	0.58	0.50
Auditory related	ETC	0.72	0.76	0.70	0.68	RF	0.69	0.65	0.65	0.56
Auditory related	GB	0.66	0.67	0.65	0.63	GB	0.69	0.66	0.65	0.56
Auditory related	Baseline	0.63	0.39	0.48	0.50	Baseline	0.70	0.49	0.58	0.50
<b>Auditory generic</b>	<b>ETC</b>	<b>0.80</b>	<b>0.83</b>	<b>0.77</b>	<b>0.75</b>	<b>ETC</b>	<b>0.77</b>	<b>0.77</b>	<b>0.74</b>	<b>0.67</b>
Auditory generic	GB	0.66	0.66	0.64	0.62	RFW	0.73	0.71	0.69	0.60
Auditory generic	Baseline	0.63	0.39	0.48	0.50	Baseline	0.70	0.49	0.58	0.50
Visual	GB	0.69	0.69	0.66	0.65	GB	0.74	0.72	0.70	0.64
Visual	ETC	0.67	0.71	0.65	0.62	ETC	0.69	0.62	0.64	0.55
Visual	Baseline	0.63	0.39	0.48	0.50	Baseline	0.70	0.49	0.58	0.50
Visual related	ETC	0.71	0.73	0.68	0.66	RF	0.76	0.74	0.70	0.61
Visual related	GB	0.69	0.69	0.68	0.67	GB	0.70	0.70	0.68	0.61
Visual related	Baseline	0.63	0.39	0.48	0.50	Baseline	0.70	0.49	0.58	0.50
Visual generic	ETC	0.71	0.72	0.69	0.66	ETC	0.68	0.66	0.65	0.57
Visual generic	GB	0.69	0.71	0.66	0.65	GB	0.69	0.66	0.65	0.59
Visual generic	Baseline	0.63	0.39	0.48	0.50	Baseline	0.70	0.49	0.58	0.50

Table 5.13: Best results of ALL (on the left) and DE (on the right) data sets for the different classifiers and subsets of features. The best two results for the F1-score and accuracy are highlighted as well as difference in the classifier ranking.



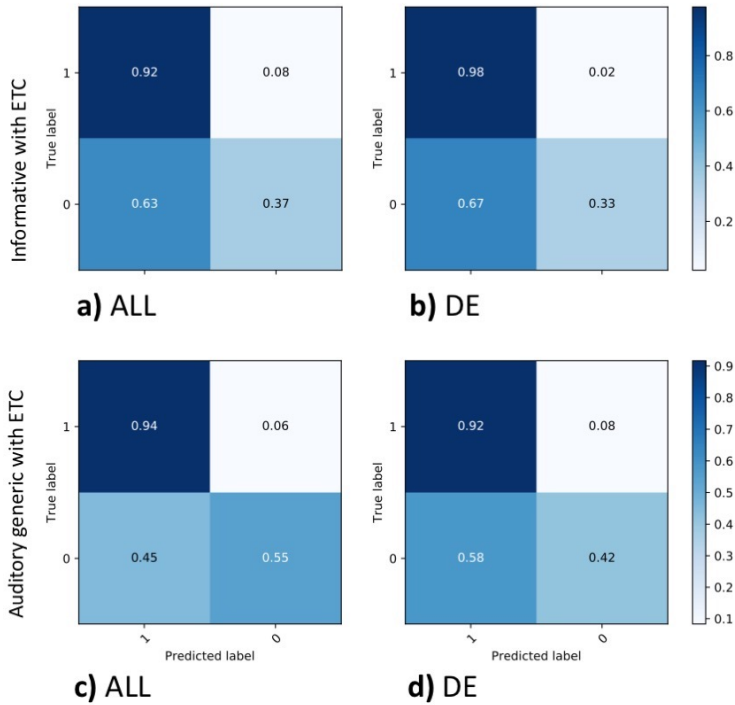


Figure 5.5: Normalised confusion matrix from the two best results (F1-score and accuracy): **a) ALL, *Informative with ETC***; **b) DE, *Informative with ETC***; **c) ALL, *Auditory generic with ETC***; and **d) DE, *Auditory generic with ETC***.

and accuracy using the feature selections *auditory generic* (0.77, 0.75) and *informative* (0.75, 0.73) and the ETC model.

For the DE data set, we achieved the best two results for the F1-score and accuracy using the feature selections *auditory generic* (0.74, 0.67) and *informative* (0.74, 0.65) and using the ETC model.

The ranking of classifiers for each selected feature is nearly the same for ALL and DE except for the *auditory* feature selection. For the auditory feature selection, ALL predicts best with RF and RFW, while DE has the best results with ETC and RF.

The prediction from *MusVis* (Chapter 4) is higher than from *DGames*. But *MusVis* has more participants for each language and therefore a bigger sample size. The auditory features are more informative in *DGames* than in *MusVis*. The *DGames* prediction for the subsets of auditory features has a higher accuracy and F1-score than for visual features.

The normalised confusion matrix (see Figure 5.5) does not show over-fitting for the best obtain results for ALL and DE.

## 5.7 Discussion

As for the participants of the game *MusVis*, we know that our participants have a high variance because of the measures of our current diagnostic tools and the wide spectrum dyslexia has. Additionally, we believe that questions regarding further diagnoses of dyslexia were not answered by the participants because the questions were too specific or the status was not known (e.g., “Does the participant have attention deficit hyperactivity disorder (ADHD) or attention deficit disorder (ADD)?”)

In the following sections we will discuss the results from the statistical analysis and the predictive models.

### 5.7.1 Statistical Comparison

The variables taken from the game *DGames* show that participants with dyslexia behave differently for nine indicators than the control group for German (R1). Differences can be reported for one indicator in the auditory game: *number of repeats*. The visual part has eight significant indicators: *total clicks*, *correct answers*, *1st click interval*, *2nd click interval*, *3rd click interval*, *time last click*, *efficiency and effect*. Due to the Bonferroni-correction, three dependent variables (duration per round, thinking time and 5th click interval) lost significance, but might regain it with more participants' for German,

Spanish and English, as dyslexia is present *across different languages and cultures* [2].

Three of our significant dependent variables have a medium effect size. This means that the experiment can explain 9% of the variances and estimates the size in the population as defined by Cohen [37]. Participants with dyslexia show a tendency to click faster, as the average is shorter for nearly all click intervals (except for the 6th click interval). Since all participants and supervisors had the same (video) instructions, we rule out the study set up as an external factor.

We now compare the statistical results from *MusVis* with the data set from *DGames* first for the auditory part and then for the visual part.

We changed the gameplay and content of *MusVis* in *DGames* because of the study results from *MusVis*. For example, to have easier gameplay, game content related to various deficits known to be connected to dyslexia and generic content were used to explore behavioral differences. Therefore, we cannot compare these results in detail with the *MusVis* auditory part, as the variables and interactions are different. But the changed auditory part of *DGames* shows with fewer participants a significant indicator, while the auditory part of *MusVis* with more participants did work for Spanish but not for German. We suspect that we can confirm the significant indicator for Spanish by enrolling more participants.

The visual game play is the same as for *MusVis* but we added generic content to the existing visual cues. As expected, significant indicators from *MuVis* are also significant for *DGames*.

We here compare the significant dependent variables from *MusVis* Spanish with the *DGames* German data set because *MusVis* German had no significant indicators. We see two possible explanations for this phenomenon. First, the content is designed to address more characteristics related to dyslexia, which should cause a stronger effect between the group of participants with and without dyslexia. Second, German *DGames* has significant indi-

cators probably because of our *repeated-measures design*, as it is more likely to measure an effect in a *repeated-measures design* [37]. This means that more game rounds in *DGames* with fewer participants produces more data and the effect caused by our experiment design is more likely to be measured. As for *MusVis*, the time to first click is different between the auditory and visual parts. We cannot compare the only added dependent variables in *DGames* 2nd to 6th click interval, time last click and effect.

### 5.7.2 Screening Differences

Our approach aims to measure dyslexia with indicators that do not involve linguistic knowledge and are generic. These indicators are probably not as strong or visible as the reading and spelling mistakes of children with dyslexia. Therefore, we consider our results for DE (highest balanced accuracy of 0.67 and highest F1-scores of 0.74) as a promising way to screen dyslexia using language-independent cues and generic content related to various dyslexia indicators (R3). Having an early detection of dyslexia before spelling or reading errors appear can have a positive impact on the child's development, as we can intervene earlier in her/his education.

The ALL data set achieved a better prediction result than DE, but the unbalanced Spanish participant group biased the prediction, something that in retrospect made sense. Due to the limited resources, only a few participants played in English and Spanish. English has a balanced group of participants, with only one participant for each group. The Spanish participant group mainly contained participants with dyslexia, with only one control participant. We assume that the model uses the unbalanced Spanish participant group as the relationship for the prediction and thus achieves higher prediction results compared to DE. An additional data collection of Spanish participants for the control group is needed to confirm our current results for ALL.

The comorbidity of dyslexia in the low-cost prediction remains a challenge. A clustering of data might help to ensure the quality of data and a better prediction. For example, it could help to cluster abilities of participants who are better for the visual or auditory game or participants who have a tendency for visual or auditory difficulties.

In our case, the subsets with auditory features have slightly better metrics than visual (see Table 5.13), although it is difficult to compare visual and auditory content (for example, based on the level of difficulty or similarity). The subsets with auditory features probably have higher prediction scores for *DGames* than *MusVis* for two reasons: (1) the *DGames* content is related to various deficits of children with dyslexia, while *MusVis* contains auditory cues primarily related to one acoustic parameter (one deficit); (2) the bigger sample size provides more information for the models (repeated measures).

Our results support the theory that a stronger effect can be measured when content is related to more indicators. As comorbidity causes additive working memory deficits [81], more indicators combined in game content could create a more prominent effect between groups (with and without dyslexia).

As described before, small data can help us to understand the data and results better. For example, we agree that using one model for different languages remains a challenge [6]. The reason that ALL performs better than DE is probably due to the unbalanced Spanish participant data, not because of the model.

We show that the prediction for dyslexia is possible with content related to various indicators in a small data set with German participants and that generic content can achieve better prediction results: ALL but also DE computed the best results with generic auditory content (ALL highest balanced accuracy of 0.75 and highest F1-scores of 0.77; DE highest balanced accuracy of 0.67 and highest F1-scores of 0.74).

# Conclusions and Future Work

---

Screening dyslexia is a growing research field, and contributions to it may help around 5% to 15% of the global population [2]. The main focus of this thesis was the design of a game based in language-independent content and the early prediction of dyslexia using machine learning with the game interaction data. The approach presented in this thesis is a starting point on how to screen for dyslexia in children without linguistic features or phonological awareness using games and a machine learning prediction using the interaction data.

In this chapter, we summarize our work and contributions. We additionally discuss open problems and future research directions on screening dyslexia with language-independent auditory and visual content. Parts of the content of this chapter were published in [114].

## 6.1 Summary

We achieved reliable results for screening dyslexia with language-independent content using machine learning with the combination

of the *design science research methodology*, *human-centred design*, experimental design, and gamification. We will reiterate our research questions (from Chapter 3) and findings (from Chapter 4 and 5). Here we summarize our answers for our three research questions:

**R1: Are there significant statistical differences between children with and without dyslexia when playing a game with auditory and visual content?**

Yes, we found significant differences between groups for different languages using auditory and visual content in the following dependent variables: The statistical analysis for *MusVis* with language-independent auditory and visual content mainly related to one indicator has seven significant variables for Spanish ( $n=153$ ): with four visual variables (total clicks, first click, hits, and efficiency) and three auditory variables (4th click, duration, and average). ALL ( $n=313$ ) has one significant visual variable (total clicks). However, we have no significant variables for German ( $n=149$ ).

The statistical analysis for *DGames* with generic language-independent auditory and visual content with various indicators has nine significant variables for German ( $n=120$ ) with eight visual variables (*total clicks, correct answers, 1st click interval, 2nd click interval, 3rd click interval, time of last click, efficiency, and effect*) and one auditory variable (*number of repeats*).

The stronger effect and the significant variables (*DGames* compared to *MusVis*) are probably caused by the content related to more characteristics of dyslexia and the *repeated-measures design*.

**R2: Is it possible to predict risk of dyslexia based on language-independent auditory and visual content using a game and machine learning for different languages?**

Yes, the prediction of dyslexia is possible with auditory and visual content using a game and machine learning for different languages, as our results in *MusVis* for different data sets show.

The machine learning classification for *MusVis* showed equal F1-scores (0.75) for German and Spanish but a higher accuracy for German (0.74) than for Spanish (0.69). However, we obtained a lower F1-score (0.65) and accuracy (0.61) for the combination of the data set (ALL). This is probably caused by the difference in the ranking of informative features for the data sets (German and Spanish), which results in features canceling each other out.

Our results support the hypothesis that dyslexia cannot be reduced to one cause but is instead a combination of auditory and visual attributes [27]. The evaluation of *MusVis* showed that people with dyslexia do not make more interaction mistakes, in spite of the fact that children with dyslexia have typically been detected by the spelling mistakes they make. Consequently, we added generic language-independent visual and auditory content related to various attributes of dyslexia and measures to a new application called *DGames*.

### **R3: Is it possible to predict risk of dyslexia based on generic language-independent visual and auditory content with various indicators using a game and machine learning?**

Yes, it is possible to predict dyslexia on generic language-independent visual and auditory content with various indicators using a game and a machine learning prediction using the interaction data. Still, the comparison of subsets of features for visual and auditory shows a slightly higher score for accuracy and F1-score when auditory content is used. Auditory subsets of features showed the best results (F1-score and accuracy) with the feature selection *auditory generic* using Extra Trees: ALL data set (0.77, 0.75,  $n = 137$ ) and DE data set (0.74, 0.67,  $n = 120$ ). The reason that ALL performs better than DE is probably due to the unbalanced Spanish participant data, not because of the model.

We addressed over-fitting using cross-validation, classifiers designed to avoid over-fitting (e.g., Random Forest with weights, Extra Trees), default parameters for classifiers, and metrics for



imbalanced data (balanced accuracy, F1-score), as well as no optimization of features within the cross-validation loop. Therefore, our results provide evidence that prediction for dyslexia with visual and auditory content is possible with the data taken from the same game. However, different models are needed for the prediction in different languages, something that in retrospect made sense.

Our approach aims to measure dyslexia with indicators that do not require linguistic knowledge and are generic. These indicators are probably not as strong or visible as the reading and spelling mistakes of children with dyslexia. Therefore, we consider our results with language-independent *auditory generic content* for DE (highest balanced accuracy of 0.67 and highest F1-scores of 0.74) using Extra Trees as a promising way to screen for dyslexia using language-independent content related to various dyslexia indicators. Early detection of dyslexia before spelling or reading errors appear can have a positive impact on a child's development, as we can intervene earlier in her/his education. Our approach can optimize resources for detecting and treating dyslexia and provide objective measures for non-professionals. Such a web game has the potential to be low-cost and easily accessible, be available for a broader audience through the Web, and make parents sensitive to the risk of dyslexia and guide them to more help (e.g., medical doctor or therapist). However, it would need at the beginning more personnel to screen all children at a young age, but could be done easily and quickly from home with the parents present. Children with dyslexia need around two years to compensate for their difficulties. Hence, this approach could help to decrease school failure, late treatment, and most importantly, to reduce suffering for children and parents.

<b>Tools</b>	<b>AGTB 5–12</b> 2012 [56, 66] German	<b>DYSL–X</b> 2013 [45, 152] Italian	<b>GC</b> 2017 [44] Italian	<b>Lexa</b> 2018 [98] English	<b>MusVis*</b> 2018 [115] German Spanish English 10–15 min.
<b>Languages</b>		n/a	endless	n/a	
<b>Duration</b>	~ 87 min.				
<b>Skills</b>					
<i>Memory general</i>	✓			✓	✓
Working memory					✓
Short-term memory	✓				✓
Auditory sequential memory					✓
Auditory phonological memory	✓			✓	✓
<b>Processing speed</b>				✓	✓
<i>Language skills</i>	✓	✓			
general					
Alphabetic awareness		✓	✓		
Phonological awareness					✓

Table 6.1: Cognitive skills tested in dyslexia screeners for pre-readers. \* *DGames* addresses the same skills as *MusVis*.

The main advantage is that this approach opens the possibility of predicting the risk of dyslexia using auditory and visual language-independent content, which has the potential to be used for pre-readers. Indeed, we aim to collect more data with younger children to validate our approach for this case. We give an overview of the cognitive skills tested compared to example applications from the related work (Chapter 2) and added the details from *MusVis* and *DGames* in Table 6.1.

To sum up, the stronger effect and significant dependent variables for auditory and visual while having a smaller sample size for *DGames* are probably caused by both the content related to more characteristics of dyslexia and the *repeated-measures design*. *MusVis* has the best F1-score (0.75) for both languages, with German using 5 features with Random Forest and Spanish using Extra Trees with 20 features related to visual and auditory content. However, predicting dyslexia for different languages with the same prediction model remains a challenge.

## 6.2 Future Directions

Future lines of research include using new languages and conducting further analysis and measures for small health data for screening dyslexia with language-independent games and using machine learning with the game interaction data. The new research field of detection presents applications accessible through the Web to screen for a person with dyslexia. Additionally, this research area opens up the opportunity to use the Web for medium-scale online experiments designed to prove hypotheses related to underrepresented target groups.

In our case, the language-independent screening needs to be tested with more languages and larger data samples to confirm our results. We already started pilot studies in Catalan and English to be able to compare the different languages and study whether

the game content is truly language-independent. For future work, we are planning to conduct a large-scale study to be able to apply a machine learning model to predict dyslexia on training, test, and validation data sets instead of only cross-validation.

Our participant calls received the attention of parents who suspected their child of having dyslexia. These parents had observed their children and mainly needed the confirmation that the indications of dyslexia were strong enough to visit a medical doctor. Our research could be implemented and tested in a product to visualize the probability of dyslexia and to advise on when to see a doctor. Therefore, in the long term, we plan to develop a game that uses language-independent auditory and visual indicators applicable for *pre-readers* and that is easily accessible from any given location with a tablet, laptop, or desktop computer. Such a game will leverage the opportunities of children with dyslexia by being able to predict their difficulties when they are pre-readers, in time to effectively intervene.

The assumption is that early detection for pre-readers is more helpful than late traditional detection tests. Early detection also means early intervention and probably a faster and more concrete knowledge of reading and writing. The challenging parts are the degree of personalization for individual learning as well as the engagement to keep the child practicing. Our results might be helpful for evaluating the progress of learning and intervention.

The analysis of data collected through online experiments with a larger number of participants will improve future detection and intervention tools, which will surely involve machine learning and other data science methods. Indeed, since we are dealing with a social problem, we should make sure we detect everyone with dyslexia and interview many more people over the years instead of not detecting people with dyslexia that may fail at school. This social problem applies to any prediction process related to health issues.

On the one hand, the possibilities for exploring dyslexia on the Web have increased in the last years. On the other hand, little

is known about the social effects: What kind of feedback does a person receive when writing on the Web with spelling mistakes? How does the (apparently) negative feedback affect a person's writing or personality? Are the time and effort a person with dyslexia has to spend on writing correctly a daily disadvantage because they cannot spend time on other things to succeed?

So far, the focus is mainly on writing and reading, and future research could take advantage of the strength that a person with dyslexia uses to compensate for the challenges of reading and writing. As a person with visual impairment trains other sensory abilities (*e.g.*, *hearing or touch sense*), a person with dyslexia must train other areas to compensate for the difficulties with texts. These compensation strategies might lead to a better understanding of dyslexia and improve the guidelines for presenting (digital) text or supporting (digital) writing. Apart from that, the disadvantage a person with dyslexia faces in their daily routine on the Web, in social media or through conversations, has an impact on each individual and on the content of the Web. When exploring user-created content or predicting diseases, a multi-modal approach is needed (*i.e.*, including different computer science fields such as HCI, IR, and ML as well as other disciplines like psychology)

The field has evolved and taken new directions in the detection and intervention through web-based applications. It is not only about assistive tools to understand what has been written or how to write correctly. Rather, the Web is now a place to explore and study dyslexia with web methods. Communication, training, and support aimed at limiting spelling mistakes are the obvious solutions for improving the writing of a person with dyslexia. Early detection is key to supporting a person with dyslexia and helping them to succeed. Although they can succeed without it, it is much more difficult. Therefore, we should not wait for a person to fail before helping. Rather, we should work to improve our methods and develop a screening tool that can be used for early detection of dyslexia.

# Bibliography

---

- [1] M. I. Ahmad and S. Shahid. Design and Evaluation of Mobile Learning Applications for Autistic Children in Pakistan. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, and M. Winckler, editors, *INTERACT*, volume 9296 of *Lecture Notes in Computer Science*, pages 436–444, Cham, 2015. Springer International Publishing.
- [2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, London, England, May 2013.
- [3] J. Arnowitz, M. Arent, and N. Berger. *Effective Prototyping for Software Makers*. Elsevier, 2007.
- [4] T. Asvestopoulou, V. Manousaki, A. Psistakis, I. Smyrnakis, V. Andreadakis, I. M. Aslanides, and M. Papadopouli. Dyslexml: Screening tool for dyslexia using machine learning, 2019.
- [5] R. Baeza-Yates. Big, small or right data: Which is the proper focus? <https://www.kdnuggets.com/2018/10/big-small-right-data.html>, 2018. [Online, accessed 22-July-2019].

- [6] A. Bandhyopadhyay, D. Dey, and R. K. Pal. *Prediction of Dyslexia using Machine Learning — A Research Travelogue*, volume 24. Springer Singapore, 2018.
- [7] P. Banerjee. *About Face 2.0: The Essentials of Interaction Design: Alan Cooper and Robert Reimann Published by John Wiley & Sons, 2003, 576 pp, ISBN 0764526413*, volume 3. Wiley Publishing, Inc., 2004.
- [8] A. A. Benasich and P. Tallal. Infant discrimination of rapid auditory cues predicts later language impairment. *Behavioural Brain Research*, 136(1):31–49, 2002.
- [9] A. A. Benasich and P. Tallal. Dyslexia risk gene relates to representation of sound in the auditory brainstem. *Developmental Cognitive Neuroscience*, 24(February):63–71, 2017.
- [10] G. Berget and A. MacFarlane. Experimental Methods in IIR. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval - CHIIR '19*, pages 93–101, New York, New York, USA, 2019. ACM Press.
- [11] S. Bertoni, A. Facchetti, S. Franceschini, C. E. Palazzi, and D. Ronzani. A Web Application for Reading and Attentional Assessments. In *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good - Goodtechs '18*, pages 142–147, New York, New York, USA, 2018.
- [12] Bildungsministerium. Rundschreiben des Bildungsministeriums: Schule und Datenschutz-Grundverordnung der EU (DSGVO) (Circular letter of the Ministry of Education: School and General Data Protection Regulation of the EU (GDPR)). [https://schulrecht-sh.de/texte/d/dgsvo\\_und\\_schule.pdf](https://schulrecht-sh.de/texte/d/dgsvo_und_schule.pdf). [Online, accessed 13-June-2019].

- [13] C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 1. Springer Science+Business Media, LLC, Singapore, 2006.
- [14] D. W. Black, J. E. Grant, and American Psychiatric Association. *DSM-5 guidebook: The essential companion to the Diagnostic and statistical manual of mental disorders, fifth edition*. American Psychiatric Association, 5th edition edition, 2016.
- [15] E. Borleffs, B. A. M. Maassen, H. Lyytinen, and F. Zwarts. Cracking the Code: The Impact of Orthographic Transparency and Morphological-Syllabic Complexity on Reading and Developmental Dyslexia. *Frontiers in Psychology*, 9(JAN):1–19, jan 2019.
- [16] H. Brau and F. Sarodnick. *Methoden der Usability Evaluation (Methods of Usability Evaluation)*. Verlag Hans Huber, Bern, 2 edition, 2006.
- [17] J. Brownlee. A Gentle Introduction to Effect Size Measures in Python. <https://machinelearningmastery.com/effect-size-measures-in-python/>. [Online, accessed 10-June-2019].
- [18] Bruno and R. M. Comprehensive Test of Phonological Processing (CTOPP). *Diagnostique*, 24(1), 2000.
- [19] K. Caine. Local standards for sample size at chi. In *CHI'16*, pages 981–992, 2016.
- [20] B. Carbol. Research Brief: Use of the Dyslexia Quest App as a Screening Tool. Technical report, Schmidt & Carbol Consulting Group, 2014.
- [21] H. W. Catts, A. McIlraith, M. S. Bridges, and D. C. Nielsen. Viewing a phonological deficit within a multifactorial model of dyslexia. *Reading and Writing*, 30(3):613–629, mar 2017.



- [22] E. Chatzidaki, M. Xenos, and C. Machaira. Let's Play a Game! Kin-LDD: A Tool for Assisting in the Diagnosis of Children with Learning Difficulties. *Multimodal Technologies and Interaction*, 3(1):16, mar 2019.
- [23] C. Coleman, N. Gregg, L. McLain, and L. W. Bellair. A Comparison of Spelling Performance Across Young Adults With and Without Dyslexia. *Assessment for Effective Intervention*, 34(2):94–105, 2008.
- [24] F. Cuetos, J. L. Ramos, and E. Ruano. PROESC. Evaluación de los procesos de escritura (Writing processes assessment). *Madrid: TEA*, 2002.
- [25] F. Cuetos, B. Rodríguez, E. Ruano, and D. Arribas. PROLEC-R: Batería de Evaluación de los Procesos Lectores, Revisada (Battery of reading processes assessment—Revised), 2007.
- [26] F. J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, mar 1964.
- [27] G. De Zubicaray and N. O. Schiller. *The Oxford handbook of neurolinguistics*. Oxford University Press, New York, NY, 2018.
- [28] S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From Game Design Elements to Gamefulness: Defining ?Gamification? *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11*, page 2425, 2011.
- [29] Deutsche Gesellschaft für Kinder und Jugendpsychiatrie Psychosomatik und Psychotherapie e.V. (DGKJP). Diagnostik und Behandlung von Kindern und Jugendlichen mit Lese- und / oder Rechtschreibstörung Registernummer 028-044

- (Diagnostics and treatment of children and adolescents with reading and / or spelling disorder register number 028-044). [https://www.awmf.org/uploads/tx\\_szleitlinien/028-0441\\_S3\\_Lese-Rechtschreibst%C3%B6rungen\\_Kinder\\_Jugendliche\\_2015-06.pdf](https://www.awmf.org/uploads/tx_szleitlinien/028-0441_S3_Lese-Rechtschreibst%C3%B6rungen_Kinder_Jugendliche_2015-06.pdf). [Online, accessed 22-Mai-2018].
- [30] T. Dietterich. Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, 27(3):326–327, sep 1995.
- [31] G. Esser, A. Wyschkon, and M. Schmidt. Was wird aus Achtjährigen mit einer Lese- und Rechtschreibstörung? Ergebnisse im Alter von 25 Jahren. (What about eight-year-olds with a reading and spelling disorder? Results at the age of 25 years.). *Zeitschrift für Klinische Psychologie und Psychotherapie*, 4:235–242, 2002.
- [32] European Commission/EACEA/Eurydice. The Structure of the European Education Systems 2018/19: Schematic Diagrams. Eurydice Facts and Figures. [https://eacea.ec.europa.eu/national-policies/eurydice/sites/eurydice/files/the\\_structure\\_of\\_the\\_european\\_education\\_systems\\_2018\\_19.pdf](https://eacea.ec.europa.eu/national-policies/eurydice/sites/eurydice/files/the_structure_of_the_european_education_systems_2018_19.pdf), 2018. [Online, accessed 22-Mai-2018].
- [33] European Union. General Data Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, repealing Directive 95/46/EC, Official Journal L 119, 4.5.: 1–88. <https://eur-lex.europa.eu/eli/reg/2016/679>, 2016.
- [34] J. J. Faraway and N. H. Augustin. When small data beats big data. *Statistics & Probability Letters*, 136:142–145, may 2018.

- [35] H. Fastl and E. Zwicker. *Psychoacoustics*. Springer Berlin Heidelberg, Berlin, Heidelberg, third edition, 2007.
- [36] A. Fawcett and R. Nicolson. *The Dyslexia Screening Test: Junior (DST-J)*. Harcourt Assessment, 2004.
- [37] A. P. Field and G. Hole. *How to design and report experiments*. SAGE Publications, London, 2003.
- [38] S. Franceschini, S. Gori, M. Ruffino, K. Pedrolli, and A. Facoetti. A Causal Link between Visual Spatial Attention and Reading Acquisition. *Current Biology*, 22(9):814–819, may 2012.
- [39] A. Frid and Z. Breznitz. An SVM based algorithm for analysis and discrimination of dyslexic readers from regular readers using ERPs. In *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, pages 1–4. IEEE, nov 2012.
- [40] N. Gaab. The Gaab Lab for Developmental Neuroscience. <http://thegaablab.com/app.html>, 2017. [Online, accessed 06-June-2019].
- [41] N. Gaab. Early Literacy Screener. <https://www.bostonearlyliteracyscreener.com/>, 2018. [Online, accessed 06-June-2019].
- [42] N. Gaab. The Screening Battery Risk Indicators Assessment. <http://thegaablab.com/imgs/PowerpointApp.pdf>, unknown. [Online, accessed 06-June-2019].
- [43] O. Gaggi, G. Galiazzo, C. Palazzi, A. Facoetti, and S. Franceschini. 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 2012 - Proceedings*, pages 1–5, Munich, Germany, 2012. IEEE.

- [44] O. Gaggi, C. E. Palazzi, M. Ciman, G. Galiazzo, S. Franceschini, M. Ruffino, S. Gori, A. Facoetti, O. Gaggi, C. E. Palazzi, G. Galiazzo, S. Franceschini, S. Gori, and A. Facoetti. Serious Games for Early Identification of Developmental Dyslexia. *Comput. Entertain. Computers in Entertainment*, 15(4), 2017.
- [45] L. Geurts, V. Vanden Abeele, V. Celis, J. Husson, L. Van den Audenaeren, L. Loyez, A. Goeleven, J. Wouters, and P. Ghesquière. DIESEL-X: A Game-Based Tool for Early Risk Detection of Dyslexia in Preschoolers. In *Describing and Studying Domain-Specific Serious Games*, pages 93–114. Springer International Publishing, Switzerland, 2015.
- [46] U. Goswami. A temporal sampling framework for developmental dyslexia. *Trends in Cognitive Sciences*, 15(1):3–10, jan 2011.
- [47] U. Goswami, L. Barnes, N. Mead, A. J. Power, and V. Leong. Prosodic Similarity Effects in Short-Term Memory in Developmental Dyslexia. *Dyslexia*, 22(4):287–304, 2016.
- [48] U. Goswami, R. Cumming, and A. Wilson. Rhythmic Perception, Music and Language: A New Theoretical Framework for Understanding and Remediating Specific Language Impairment Background to the Project. Technical report, University of Cambridge, 2016.
- [49] U. Goswami, N. Mead, T. Fosker, M. Huss, L. Barnes, and V. Leong. Impaired perception of syllable stress in children with dyslexia: A longitudinal study. *Journal of Memory and Language*, 69(1):1–17, 2013.
- [50] GRAPHO GROUP. GraphoGame — Home. <https://www.graphogame.com/>. [Online, accessed 19-June-2019].

- [51] M. Grund, C. L. Naumann, and G. Haug. *Diagnostischer Rechtschreibtest für 5. Klassen: DRT 5 (Diagnostic spelling test for fifth grade: DRT 5)*. Deutsche Schultests. Beltz Test, Göttingen, 2., aktual edition, 2004.
- [52] T. K. Guttorm, P. H. T. Leppänen, A. Tolvanen, and H. Lyytinen. Event-related potentials in newborns with and without familial risk for dyslexia: principal component analysis reveals differences between the groups. *Journal of Neural Transmission*, 110(9):1059–1074, sep 2003.
- [53] J. A. Hämäläinen, H. K. Salminen, and P. H. T. Leppänen. Basic Auditory Processing Deficits in Dyslexia: Systematic Review of the Behavioral and Event-Related Potential/ Field Evidence. *Journal of Learning Disabilities*, 46(5):413–427, 2013.
- [54] J. Hamari, J. Koivisto, and H. Sarsa. Does Gamification Work? – A Literature Review of Empirical Studies on Gamification. In *2014 47th Hawaii International Conference on System Sciences*, pages 3025–3034. IEEE, jan 2014.
- [55] M. Hasselhorn, R. Schumann-Hengsteler, J. Gronauer, D. Grube, C. Mähler, I. Schmid, K. Seitz-Stein, and C. Zoelch. AGTB 5-12 — Arbeitsgedächtnistestbatterie für Kinder von 5 bis 12 Jahren (AGTB 5-12 — Working memory tasks to test children at the age of 5 till 12). <https://www.testzentrale.de/shop/arbeitsgedaechtnistestbatterie-fuer-kinder-von-5-bis-12-jahren.html>, 2012. [Online, accessed 04-June-2019].
- [56] M. Hasselhorn and C. Zoelch. *Funktionsdiagnostik des Arbeitsgedächtnisses (Functional diagnostics of the working memory)*. Hogrefe Verlag, Göttingen, 2012.

- [57] T. Helland and R. Kaasa. Dyslexia in English as a second language. *Dyslexia*, 11(1):41–60, feb 2005.
- [58] A. Hevner, S. T. March, J. Park, and S. Ram. Design Science in Information Systems Research. *MIS Quarterly*, 28(1):75, 2004.
- [59] A. Hinderks, M. Schrepp, M. Rauschenberger, S. Olschner, and J. Thomaschewski. Konstruktion eines Fragebogens für jugendliche Personen zur Messung der User Experience. (Construction of a questionnaire for young people to measure user experience.). In *Usability Professionals Konferenz 2012*, pages 78–83. German UPA e.V., Stuttgart, 2012.
- [60] B. Hornsby and P. Cox. *Overcoming dyslexia : a straightforward guide for families and teachers*. Vermilion, London, 1996.
- [61] R. R. Huffman, A. Roesler, and B. M. Moon. What is design in the context of human-centered computing? *IEEE Intelligent Systems*, 19(4):89–95, 2004.
- [62] K. Huotari and J. Hamari. Defining gamification. In *Proceeding of the 16th International Academic MindTrek Conference on - MindTrek '12*, MindTrek '12, page 17, New York, NY, USA, 2012. ACM.
- [63] M. Huss, J. P. Verney, T. Fosker, N. Mead, and U. Goswami. Music, rhythm, rise time perception and developmental dyslexia: Perception of musical meter predicts reading and phonology. *Cortex*, 47(6):674–689, jun 2011.
- [64] IDC Worldwide. Shipment forecast of tablets, laptops and desktop PCs worldwide from 2010 to 2019 (in million units). *Statista*, 2020:2020, 2016.

- [65] International Dyslexia Association. Frequently Asked Questions, 2019.
- [66] D. Irblich, S. K. Diakonie, and G. Renner. AGTB 5-12 — Arbeitsgedächtnistestbatterie für Kinder von 5 bis 12 Jahren (AGTB 5-12 — Working memory tasks to test children at the age of 5 till 12). *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 62(5):1–389, 2013.
- [67] ISO/TC 159/SC 4 Ergonomics of human-system interaction. Part 210: Human- centred design for interactive systems. In *Ergonomics of human-system interaction*, volume 1, page 32. International Organization for Standardization (ISO), Brussels, 2010.
- [68] ISO/TC 159/SC 4 Ergonomics of human-system interaction. ISO 9241-11, Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts, 2018.
- [69] A. Jain and D. Zongker. Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.
- [70] H. Jansen, G. Mannhaupt, H. Marx, and H. Skowronek. *BISC - Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten (Early detection with the Bielefelder screening of reading and spelling difficulties)*. Hogrefe, Verlag für Psychologie, Göttingen, 2002.
- [71] D. J. Johnson. Persistent auditory disorders in young dyslexic adults. *Bulletin of the Orton Society*, 30(1):268–276, jan 1980.
- [72] L. M. Justice, W.-Y. Ahn, and J. A. R. Logan. Identifying Children With Clinical Language Disorder: An Application of

Machine-Learning Classification. *Journal of Learning Disabilities*, pages 1–15, 2019.

- [73] I. Kecskes and T. Papp. *Foreign Language and Mother Tongue*. Psychology Press, New York, 1 edition, jun 2000.
- [74] B. Kitchenham and P. Brereton. A systematic review of systematic review process research in software engineering. *Information and Software Technology*, 55(12):2049–2075, dec 2013.
- [75] F. Kyle, J. Kujala, U. Richardson, H. Lyytinen, and U. Goswami. Assessing the Effectiveness of Two Theoretically Motivated Computer-Assisted Reading Interventions in the United Kingdom: GG Rime and GG Phoneme. *Reading Research Quarterly*, 48(1):61–76, 2013.
- [76] F. Laamarti, M. Eid, and A. El Saddik. Review Article An Overview of Serious Games. *International Journal of Computer Games Technology*, 2014(15 October 2014):1–15, 2014.
- [77] A. Le Floch and G. Ropars. Left–right asymmetry of the Maxwell spot centroids in adults without and with dyslexia. *Proceedings of the Royal Society B: Biological Sciences*, 284(1865):20171380, oct 2017.
- [78] V. Levenshtein. Binary codes capable of correctingspurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17, 1965.
- [79] Lexercise. Dyslexia Test - Online from Lexercise. <http://www.lexercise.com/tests/dyslexia-test>, 2016. [Online; accessed 18-September-2017].
- [80] H. Lyytinen, J. Erskine, J. Hämäläinen, M. Torppa, and M. Ronimus. Dyslexia - Early Identification and Prevention:



Highlights from the Jyväskylä Longitudinal Study of Dyslexia. *Current Developmental Disorders Reports*, 2(4):330–338, dec 2015.

- [81] C. Maehler and K. Schuchardt. Working memory in children with specific learning disorders and/or attention deficits. *Learning and Individual Differences*, 49:341–347, 2016.
- [82] C. Männel, G. Schaadt, F. K. Illner, E. van der Meer, and A. D. Friederici. Phonological abilities in literacy-impaired children: Brain potentials reveal deficient phoneme discrimination, but intact prosodic processing. *Developmental Cognitive Neuroscience*, 23:14–25, 2016.
- [83] G. Mioni, A. Capodieci, V. Biffi, F. Porcelli, and C. Cornoldi. Difficulties of children with symptoms of attention-deficit/hyperactivity disorder in processing temporal information concerning everyday life events. *Journal of Experimental Child Psychology*, 182:86–101, jun 2019.
- [84] A. Mora, D. Riera, C. Gonzalez, and J. Arnedo-Moreno. A Literature Review of Gamification Design Frameworks. In *VS-Games 2015 - 7th International Conference on Games and Virtual Worlds for Serious Applications*, 2015.
- [85] C. Moritz, S. Yampolsky, G. Papadelis, J. Thomson, and M. Wolf. Links between early rhythm skills, musical training, and phonological awareness. *Reading and Writing*, 26(5):739–769, 2013.
- [86] Nesy. Dyslexia screening - Nesy UK. <https://www.nesy.com/uk/product/dyslexia-screening/>, 2011. [Online; accessed 18-September-2017].
- [87] R. I. Nicolson and A. J. Fawcett. Comparison of deficits in cognitive and motor skills among children with dyslexia. *Annals of Dyslexia*, 44(1):147–164, 1994.

- [88] Niedersächsisches Vorschrifteninformationssystem. VORIS Kultusministerium | 25b - 81402 | Verwaltungsvorschrift (Niedersachsen) | Umfragen und Erhebungen in Schulen | i. d. F. v. 01.12.2015 | gültig ab 01.12.2015 | gültig bis 31.12.2019 (VORIS Ministry of Education and Cultural Affairs | 25b - 81402 | Administrative regulations (Lower Saxony) | Polls and surveys in schools | in the version of 01.12.2015 | valid from 01.12.2015 | valid until 31.12.2019). <http://www.nds-voris.de/jportal/?quelle=jlink&query=VVND-224100-MK-20140101-SF&psml=bsvorisprod.psml&max=true>. [Online, accessed 13-June-2019].
- [89] J. Nielsen. Why You Only Need to Test with 5 Users. *Jakob Niensens Alertbox*, 19(September 23):1–4, 2000. [Online, accessed 11-July-2019].
- [90] J. Nijakowska. *Dyslexia in the foreign language classroom*. Multilingual Matters, 2010.
- [91] D. Norman. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, New York, nov 2013.
- [92] K. Overy. Dyslexia, Temporal Processing and Music: The Potential of Music as an Early Learning Aid for Dyslexic Children. *Psychology of Music*, 28(2):218–229, oct 2000.
- [93] O. Ozernov-Palchik, M. Gonzalez, L. Hillyer, J. Diefenbach, J. Gabrieli, and N. Gaab. Copy of Early Literacy Assessments. [https://docs.google.com/spreadsheets/d/16m40o49LZ\\_9wZI9VPaxhHF1ATvhSM1mm-0oGr48jFfo/edit#gid=734901246](https://docs.google.com/spreadsheets/d/16m40o49LZ_9wZI9VPaxhHF1ATvhSM1mm-0oGr48jFfo/edit#gid=734901246), 2019. [Online, accessed 06-June-2019].
- [94] E. Paulesu, L. Danelli, and M. Berlingeri. Reading the dyslexic brain: multiple dysfunctional routes revealed by a

new meta-analysis of PET and fMRI activation studies. *Frontiers in human neuroscience*, 8:830, 2014.

- [95] P. M. Paz-Alonso, M. Oliver, G. Lerma-Usabiaga, C. Caballero-Gaudes, I. Quiñones, P. Suárez-Coalla, J. A. Duñabeitia, F. Cuetos, and M. Carreiras. Neural correlates of phonological, orthographic and semantic reading processing in dyslexia. *NeuroImage. Clinical*, 20:433–447, 2018.
- [96] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(8):45–78, 2007.
- [97] Y. Petscher, H. Fien, C. Stanley, B. Gearin, J. M. Fletcher, Y. Petscher, C. Stanley, B. Gearin, N. Gaab, J. M. Fletcher, and &. Johnson. Screening for Dyslexia. Technical report, National Center on Improving Literacy, 2019. [Online, accessed 06-June-2019].
- [98] A. Poole, F. Zulkernine, and C. Aylward. Lexa: A tool for detecting dyslexia through auditory processing. *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings*, 2018-January:1–5, 2018.
- [99] R. F. Port. Meter and speech. *Journal of Phonetics*, 31:599–611, 2003.
- [100] Precision Learning Center. App Development. <http://www.precisionlearningcenter.org/research-app.html>. [Online, accessed 06-June-2019].
- [101] Precision Learning Center. School Partnership. <http://www.precisionlearningcenter.org/school-partnership.html>. [Online, accessed 06-June-2019].

- [102] A. Protopapas, A. Fakou, S. Drakopoulou, C. Skaloumbakas, and A. Mouzaki. What do spelling errors tell us? Classification and analysis of errors made by Greek schoolchildren with and without dyslexia. *Reading and Writing*, 26(5):615–646, may 2013.
- [103] M. Ptok, K. Berendes, S. Gottal, B. Grabherr, J. Schneeberg, and M. Wittler. Lese-Rechtschreib-Störung: Die Bedeutung der phonologischen Informationsverarbeitung für den Schriftspracherwerb (Dyslexia: The meaning of phonological information processing for the acquisition of written language). *Hno*, 55(9):737–748, 2007.
- [104] K. Pugh and L. Verhoeven. Introduction to This Special Issue: Dyslexia Across Languages and Writing Systems. *Scientific Studies of Reading*, 22(1):1–6, 2018.
- [105] M. Rauschenberger. Entwicklung von Designentwürfen zur Unterstützung von Hafenmanövern für Lotsen mittels der Hierarchischen Aufgabenanalyse. (Deriving designs in harbour manoeuvre for harbour pilots with the hierarchical task analysis.), 2015.
- [106] M. Rauschenberger. DysMusic: Detecting Dyslexia by Web-based Games with Music Elements. In *The Web for All conference Addressing information barriers – W4A’16*, Montreal, Canada, 2016. ACM Press.
- [107] M. Rauschenberger, S. Füchsel, L. Rello, C. Bayarri, and A. Görriz. A game to target the spelling of German children with dyslexia. In *Proceedings of the 17th international ACM SIGACCESS conference on Computers & accessibility - ASSETS ’15*, Lisbon, Portugal, 2015.
- [108] M. Rauschenberger, S. Füchsel, L. Rello, C. Bayarri, and J. Thomaschewski. Exercises for German-speaking children

with dyslexia. In *Human-Computer Interaction–INTERACT 2015*, pages 445–452, Bamberg, Germany, 2015.

- [109] M. Rauschenberger, S. Füchsel, L. Rello, and J. Thomaschewski. DysList German resource: A language resource of German errors written by children with dyslexia. <https://zenodo.org/record/809801#.X0VWRFMzYWo>, 2017. [Online, accessed 06-June-2019].
- [110] M. Rauschenberger, A. Hinderks, and J. Thomaschewski. Benutzererlebnis bei Unternehmenssoftware: Ein Praxisbericht über die Umsetzung attraktiver Unternehmenssoftware. (Enterprise Software User Experience: A real-world report on how enterprise software can be made attractive.). In *Usability Professionals Konferenz 2011*, volume 1, pages 154–158. German UPA e.V., Stuttgart, 2011.
- [111] M. Rauschenberger, C. Lins, N. Rousselle, S. Fudickar, and A. Hain. A Tablet Puzzle to Target Dyslexia Screening in Pre-Readers. In *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good - GOODTECHS*, pages 155–159, Valencia, 2019.
- [112] M. Rauschenberger, S. Olschner, M. P. Cota, M. Schrepp, and J. Thomaschewski. Measurement of user experience: A Spanish Language Version of the User Experience Questionnaire (UEQ). In A. Rocha, J. A. CalvoManzano, L. P. Reis, and M. P. Cota, editors, *Sistemas Y Tecnologias De Informacion*, pages 471–476, Madrid, Spain, 2012.
- [113] M. Rauschenberger, L. Rello, and R. Baeza-Yates. A Tablet Game to Target Dyslexia Screening in Pre-readers. In *MobileHCI'18*, pages 306–312, Barcelona, 2018. ACM Press.
- [114] M. Rauschenberger, L. Rello, and R. Baeza-Yates. Technologies for Dyslexia. In Y. Yesilada and S. Harper, editors,

*Web Accessibility Book*, volume 1, pages 603–627. Springer-Verlag London, London, 2 edition, 2019.

- [115] M. Rauschenberger, L. Rello, R. Baeza-Yates, and J. P. Bigham. Towards language independent detection of dyslexia with a web-based game. In *W4A '18: The Internet of Accessible Things*, pages 4–6, Lyon, France, 2018. ACM.
- [116] M. Rauschenberger, L. Rello, R. Baeza-Yates, E. Gomez, and J. P. Bigham. Supplement:DysMusicMusicalElements: Towards the Prediction of Dyslexia by a Web-based Game with Musical Elements. <https://doi.org/10.5281/zenodo.809783>, June 2017. [Online, accessed 09-June-2019].
- [117] M. Rauschenberger, L. Rello, R. Baeza-Yates, E. Gomez, and J. P. Bigham. Towards the Prediction of Dyslexia by a Web-based Game with Musical Elements. In *The Web for All conference Addressing information barriers – W4A'17*, pages 4–7, Perth, Western Australia, 2017. ACM Press.
- [118] M. Rauschenberger, L. Rello, S. Füchsel, and J. Thomaschewski. A language resource of German errors written by children with dyslexia. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).
- [119] M. Rauschenberger, M. Schrepp, M. P. Cota, S. Olschner, and J. Thomaschewski. Efficient Measurement of the User Experience of Interactive Products. How to use the User Experience Questionnaire (UEQ). Example: Spanish Language. *International Journal of Artificial Intelligence and Interactive Multimedia (IJIMAI)*, 2(1):39–45, 2013.
- [120] M. Rauschenberger, M. Schrepp, and J. Thomaschewski. User Experience mit Fragebögen messen–Durchführung

und Auswertung am Beispiel des UEQ (Measuring User Experience with Questionnaires—Execution and Evaluation using the Example of the UEQ). In *Usability Professionals Konferenz 2013*, pages 72–76, 2013.

- [121] M. Rauschenberger, A. Willems, M. Ternieden, and J. Thomaschewski. Gamification and learning environments — protocol for the systematic literature review (protocol v1.3). <https://doi.org/10.5281/zenodo.3257277>, May 2018. [Online, accessed 04.-Juli-2019].
- [122] M. Rauschenberger, A. Willems, M. Ternieden, and J. Thomaschewski. Towards the use of gamification frameworks in learning environments. *Journal of Interactive Learning Research*, 30(2), 2019.
- [123] L. Rello. *DysWebxia: A Text Accessibility Model for People with Dyslexia* Luz Rello. PhD thesis, Universtat Pompeu Fabra, 2014.
- [124] L. Rello and R. Baeza-Yates. Good fonts for dyslexia. *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, page 14, 2013.
- [125] L. Rello, R. Baeza-Yates, A. Ali, J. P. Bigham, and M. Serra. Predicting risk of dyslexia with an online gamified test. *arXiv preprint arXiv:1906.03168*, V.1:1–13, jun 2019.
- [126] L. Rello, R. Baeza-Yates, and J. Llisterri. A resource of errors written in Spanish by people with dyslexia and its linguistic, phonetic and visual analysis. *Language Resources and Evaluation*, 51(2):1–30, feb 2016.
- [127] L. Rello, M. Ballesteros, A. Ali, M. Serra, D. Alarcón, and J. P. Bigham. Dyetective: Diagnosing Risk of Dyslexia with a Game. In *Pervasive Health 2016*, pages 89–96, Cancun, Mexico, may 2016. ACM Press.

- [128] L. Rello, C. Bayarri, Y. Otal, and M. Pielot. A computer-based method to improve the spelling of children with dyslexia. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility - ASSETS '14*, ASSETS '14, pages 153–160, New York, USA, 2014. ACM Press.
- [129] L. Rello, E. Romero, M. Rauschenberger, A. Ali, K. Williams, J. P. Bigham, and N. C. White. Screening Dyslexia for English Using HCI Measures and Machine Learning. In *Proceedings of the 2018 International Conference on Digital Health - DH '18*, pages 80–84, New York, New York, USA, 2018. ACM Press.
- [130] H. Remschmidt and S. von Aster. *Kinder- und Jugendpsychiatrie: Eine praktische Einführung (Child and Adolescent Psychiatry: A Practical Introduction)*. Thieme, Stuttgart [u.a.], 4., edition edition, 2005.
- [131] J. W. Rice. The Gamification of Learning and Instruction. *International Journal of Gaming and Computer-Mediated Simulations*, 4(4):81–83, 2013.
- [132] A. D. Ritzhaupt, N. D. Poling, C. A. Frey, and M. C. Johnson. A Synthesis on Digital Games in Education: What the Research Literature Says from 2000 to 2010. *Jl. of Interactive Learning Research*, 25(2):263–282, 2014.
- [133] E. J. Rolka and M. J. Silverman. A systematic review of music and dyslexia. *Arts in Psychotherapy*, 46:24–32, 2015.
- [134] R. Rouse. *Game Design: Theory and Practice, Second Edition: Theory and Practice, Second Edition*. Wordware Publishing, Inc., 2004.
- [135] C. Rowland and M. Charlier. *User Experience Design for the Internet of Things*. O'Reilly Media, Inc., 2015.



- [136] B. Sawyer. The "Serious Games" Landscape. Presented at the Instructional & Research Technology Symposium for Arts, Humanities and Social Sciences, Camden, USA., 2007.
- [137] J. Schell. Design Outside the box: DICE 2010. Dice Summit, 2010.
- [138] G. Schulte-Körne. Diagnostik und Therapie der Lese-Rechtschreib-Störung (The prevention, diagnosis, and treatment of dyslexia). *Deutsches Ärzteblatt international*, 107(41):718–727, 2010.
- [139] G. Schulte-Körne, W. Deimel, M. Jungermann, and H. Remschmidt. Nachuntersuchung einer Stichprobe von lese-rechtschreibgestörten Kindern im Erwachsenenalter (Follow-up examination of a sample of children with reading and spelling difficulties in adulthood). *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, 31(4):267–276, nov 2003.
- [140] G. Schulte-Körne, W. Deimel, K. Müller, C. Gutenbrunner, and H. Remschmidt. Familial Aggregation of Spelling Disability. *Journal of Child Psychology and Psychiatry*, 37(7):817–822, oct 1996.
- [141] Scikit-learn. 3.1. Cross-validation: evaluating estimator performance. [https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html), 2019. [Online, accessed 17-June-2019].
- [142] Scikit-learn. 3.3. Model evaluation: quantifying the quality of predictions. [https://scikit-learn.org/stable/modules/model\\_evaluation.html#scoring-parameter](https://scikit-learn.org/stable/modules/model_evaluation.html#scoring-parameter), 2019. [Online, accessed 23-July-2019].

- [143] Scikit-learn Developers. Scikit-learn Documentation. <https://scikit-learn.org/stable/documentation.html>. [Online, accessed 20-June-2019].
- [144] K. Seaborn and D. I. Fels. Gamification in theory and action: A survey. *International Journal of Human Computer Studies*, 74:14–31, 2015.
- [145] P. H. K. Seymour, M. Aro, and J. M. Erskine. Foundation literacy acquisition in European orthographies. *British journal of psychology (London, England : 1953)*, 94(Pt 2):143–74, may 2003.
- [146] D. Sheskin. Handbook of Parametric and Nonparametric Statistical Procedures. *Boca Raton: CRC*, page 972, 2000.
- [147] H. A. Simon. *The sciences of the artificial, (third edition)*, volume 3. MIT Press, 1997.
- [148] C. Steinbrink and T. Lachmann. *Lese-Rechtschreibstörung (Dyslexia)*. Springer Berlin Heidelberg, 2014.
- [149] P. Tallal. Improving language and literacy is a matter of time. *Nature reviews. Neuroscience*, 5(9):721–728, 2004.
- [150] P. Tamboer, H. C. M. Vorst, S. Ghebreab, and H. S. Scholte. Machine learning and dyslexia: Classification of individual structural neuro-imaging scans of students with and without dyslexia. *NeuroImage. Clinical*, 11:508–514, 2016.
- [151] UCSF Dyslexia Center. AppRISE for Preschoolers to Kindergartners. <https://dyslexia.ucsf.edu/content/iscreener-preschoolers-kindergartners>. [Online, accessed 06-June-2019].
- [152] L. Van den Audenaeren, V. Celis, V. Vanden Abeele, L. Geurts, J. Husson, P. Ghesquière, J. Wouters, L. Loyez,

- and A. Goeleven. DYSL-X: Design of a tablet game for early risk detection of dyslexia in preschoolers. In *Games for Health*, pages 257–266. Springer Fachmedien Wiesbaden, Wiesbaden, 2013.
- [153] V. Vanden Abeele, J. Wouters, P. Ghesquière, A. Goeleven, and L. Geurts. Game-based Assessment of Psycho-acoustic Thresholds. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*, pages 331–341, New York, New York, USA, 2015. ACM Press.
- [154] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, feb 2006.
- [155] F. R. Vellutino, J. M. Fletcher, M. J. Snowling, and D. M. Scanlon. Specific reading disability (dyslexia): what have we learned in the past four decades? *Journal of Child Psychology and Psychiatry*, 45(1):2–40, jan 2004.
- [156] T. R. Vidyasagar and K. Pammer. Dyslexia: a deficit in visuo-spatial attention, not in phonological processing. *Trends in Cognitive Sciences*, 14(2):57–63, 2010.
- [157] T. Wallbaum, M. Rauschenberger, J. Timmermann, W. Heuten, and S. C. Boll. Exploring Social Awareness. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–10, New York, New York, USA, 2018. ACM Press.
- [158] J. G. Walls, G. R. Widmeyer, and O. A. El Sawy. Building an information system design theory for vigilant EIS. *Information Systems Research*, 3(1):36–59, 1992.
- [159] J. M. Ward-Lonergan and J. K. Duthie. The State of Dyslexia: Recent Legislation and Guidelines for Serving School-Age

Children and Adolescents With Dyslexia. *Language Speech and Hearing Services in Schools*, 49(4):810, 2018.

- [160] K. Werbach and D. Hunter. *For the win : how game thinking can revolutionize your business*. Wharton, 2012.
- [161] K. Werbach and D. Hunter. *The gamification toolkit : dynamics, mechanics, and components for the win*. Wharton Digital Press, Philadelphia :, 2015.
- [162] Wikipedia. Memory (Spiel) (Memory Game). [https://de.wikipedia.org/wiki/Memory\\_\(Spiel\)](https://de.wikipedia.org/wiki/Memory_(Spiel)), 2019. [Online, accessed 22-Mai-2018].
- [163] World Health Organization. *International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD)*. World Health Organization, 2010.
- [164] World Health Organization. *International Classification of Diseases 11th Revision*. World Health Organization, 2019.
- [165] C. J. Yuskaitis, M. Parviz, P. Loui, C. Y. Wan, and P. L. Pearl. Neural Mechanisms Underlying Musical Pitch Perception and Clinical Applications Including Developmental Dyslexia. *Current neurology and neuroscience reports*, 15(8):51, 2015.
- [166] G. Zichermann and C. Cunningham. *Gamification by Design*. O'Reilly books, Canada, 2011.
- [167] J. C. Ziegler, C. Perry, A. Ma-Wyatt, D. Ladner, and G. Schulte-Körne. Developmental dyslexia in different languages: language-specific or universal? *Journal of experimental child psychology*, 86(3):169–93, nov 2003.



# Appendix

---

## A.1 Towards the Use of Gamification

Because motivation is a key factor for successful learning and engaging children is important, gamification is a suitable mechanism to design an environment such as a product, a prototype or an application for different use cases. Frameworks have already been developed to transfer the concept of gamification to learning environments. This is especially achieved by integrating game elements into applications. Game elements are classified in different abstract levels to realize gamification in an application. For example, the highest abstract level, *dynamics*, could be *emotions* or *progression*, while the middle abstract level, *mechanics*, ,e.g., the abstract level dynamic with emotions (dynamics) is referencing to *rewards* or *feedback*. Nevertheless, the existing frameworks and game elements have not yet been compared to make the benefits visible or to clarify preferences of game elements for educational environments.

However, there are papers exploring aspects related to gamification in literature (e.g., digital games in Education [132]). There are also papers focusing on gamification frameworks in learning environments without explicitly performing a SLR. For example,

Seaborn and Fels [144] provide a comprehensive overview of applied gamification while focusing on the gap between theory and practice. It also covers gamification in education.

Theoretical foundations and frameworks are examined, yet there is no overview about papers that provide working frameworks for gamification in learning environments. Mora, Riera, Gonzalez, and Arnedo-Moreno [84] focus on frameworks and methods, but there is no conclusion as to whether frameworks are suited for educational use. Hamari, Koivisto, and Sarsa [54] provide a literature review of peer-reviewed empirical studies on gamification. It focuses on the effects, results, motivational affordances, and psychological and behavioral outcomes of the use of gamification. In conclusion, the papers mentioned earlier offer an overview of studies that focus on the outcomes, the frameworks and methods, and applications of gamification. The contents of this Appendix A.1 were published in [121, 122].

With a quick *systematic literature review* (SLR) [74] in two well-known databases in the field of computer science, the hits were reduced by applying different quality criteria, resulting in ten articles published before June 2016. All frameworks were designed during the last three years. The search protocol is available at <https://github.com/Rauschii/slrgamification2018> [121].

The increasing number of academic publications since 2014 for gamification might indicate the establishment of frameworks and game elements. The use cases for the game elements are very diverse and have not yet been evaluated. Since we did not find other educational environments with the applied SLR, no standard approach for the usage of game elements can be derived.

Therefore, the game elements are analyzed by their quantity to understand which game elements are used in regard to the different abstraction levels defined by Werbach and Hunter [160]. The game element hierarchy of *dynamics, mechanics and components*, as explained in Chapter 2, is used to be able to compare the extracted game elements from the SLR results. The differentiation

<b>Dynamics</b>	<b>Mechanics</b>	<b>Components</b>
<b>Emotions (41)</b>	<b>Rewards (31)</b>	<b>Badges (7), Points (6), Achievement (5), Leaderboard (5), Reward (3), Altruism (1), Certificates (1), Gifts (1), Score (1), Titles (1)</b>
	Resource acquisition (2)	Collectible Cards (1), Self Expression (1)
	Chance (1)	Random Questions (1)
	Win states (2)	Win States (2)
	Others (5)	Avatars (2), Emotional Experiences (1), Emotions (1), Players (1)
<b>Progression (35)</b>	<b>Feedback (13)</b>	Instant Feedback (4), Progress Bar (4), Visible Status (2), Accrual Grading (1), Skill Tree (1), Status (1)
	<b>Challenges (12)</b>	Challenges (4), Quests (3), Quizzes (2), Boss Fights (1), Goals (1), Increase of Difficulty (1)
	Rewards (5)	Levels (5)
	Others (5)	Progression (5)
Relationships (19)	Cooperation (9)	Cooperation (4), Collaboration (2), Teams (3)
	Competition (5)	Competition (5)
	Transactions (3)	Sharing (1), Virtual Goods (2)
	Others (2)	Social Networks (1), Social Recognition (1)
Narrative (9)	Rewards (3)	Unlocking Content (2), Unlock Item (1)
	Others (6)	Narrative (3), Environment (1), Interaction (1), Story (1)
Choice (8)	Feedback (1)	Hint/Tip (1)
	Rewards (1)	Quantifiable Outcomes (1)
	Others (6)	Freedom to Fail (2), Freedom of Choice (1), Negotiable Consequences (1), Possibility to Fail (1), Tutorials (1)
Constraints (3)	Challenges (3)	Rules (2), Time Limit (1)

Figure A.1: Overview of the abstraction levels: 115 Dynamics, Mechanics and Components.



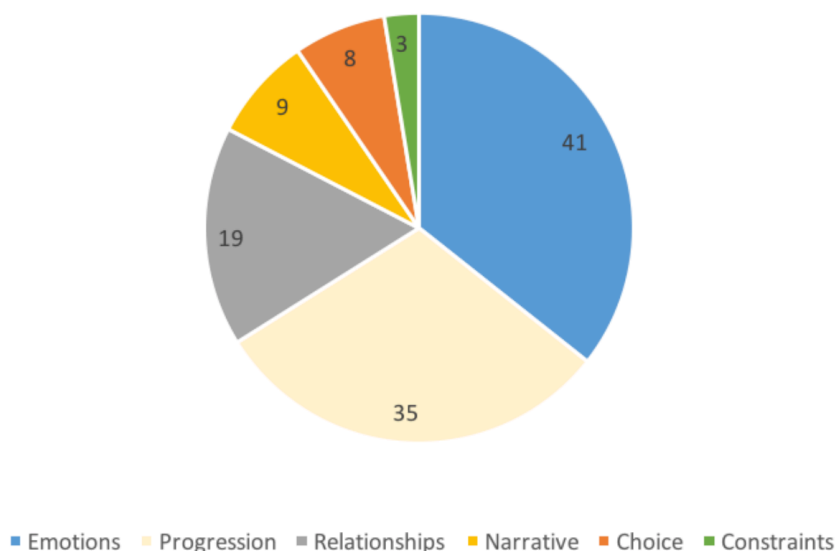


Figure A.2: Distribution of dynamics ( $n = 115$ ).

between three abstraction levels consequently allows a comparison not only of the game elements but also of the respective level in the hierarchy (see Figure A.1). The highest level dynamics is extended by choice to also include the choice of a player [161].

In total, 115 game elements are extracted from the articles and compared. Over half (76/115) of the components are associated with the *dynamics* of *emotions* and *progression* (see Figure A.2). This finding indicates that the desired increase in motivation is achieved through emotional engagement and the visualization of the students' progress.

The abstract level mechanics contains 19 different mechanics (see Figure A.1, middle column) and by far the most frequent mechanic is rewards (31/115, emotions). The mechanics feedback (13/115, progression) and challenges (12/115, progression) also

have high frequencies compared to the remaining 16 mechanics, which are mostly mentioned less than half as often. In total, 57 components (see Figure A.1, right column) are extracted. Over half of the components (30/57) are mentioned only once over all extracted articles. Furthermore, if the components that are mentioned twice (12/57) are added, over all extracted articles only 26% of the components are mentioned more than twice. Because of that, we report the components by their appearances over all dynamics and mechanics (see Figure A.1).

These new components have been introduced in the last three years and belong mainly to the dynamics emotions (11/30). Presented in order of frequency distribution, the remaining components are: levels (5), progression (5), achievement (5), leader board (5), competition (5), instant feedback (4), progress bar (4), challenges (4), cooperation (4), quests (3), reward (3), teams (3), and narrative (3).

Our results confirm the acceptance of the term gamification for educational environments. But as the variety of game elements is still widespread, the only common ground seems to be the definition of the term gamification made by Deterding [28]: Gamification applies game design elements in non-game contexts.

These findings are complementary to Mora *et al.* [84] in terms of future publications and the analysis of game elements. Also, we can confirm their previous observation on which game elements are applied.

To summarize these findings, over half of the extracted components are mentioned only once or twice. Consequently, no guidelines for the use of components can be derived to make recommendations for educational environments. The frequency distribution indicates preferences (see Figure A.1) for the different dynamics and mechanics as presented earlier. Most game elements used for educational environments are in the dynamics *emotions* and *progression*. The small number of hits for the search phrases indicates that more research on different aspects of gamification is necessary

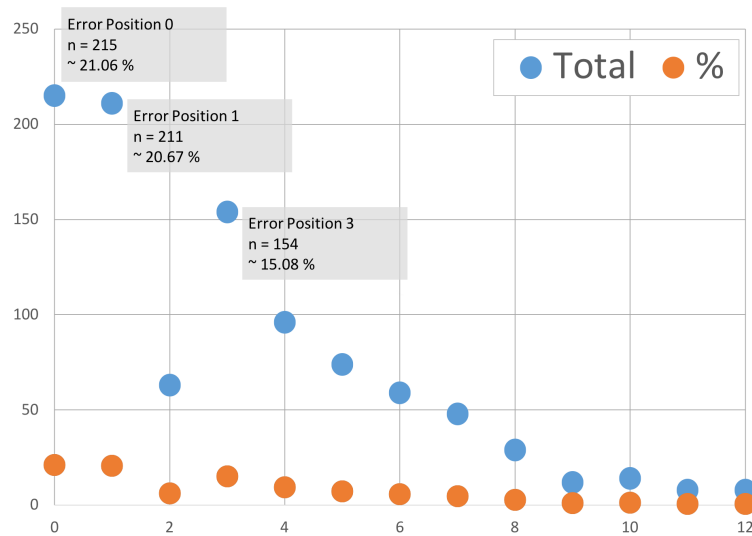


Figure A.3: Distribution of the error position for the German error resource [109] for position 0 to 12.

to provide a more representative data set for a full-fledged literature review, which is not the scope of this thesis. The results indicate preferences for the use of game elements in the different abstract levels.

## A.2 Analysis of German Error Words

Here we present an analysis of the collected error words from [109, 118], which are 1,021 error words from children with dyslexia. The analysis of the errors for German as well as Spanish and English [126] describe the auditory attributes in Table 5.1. Now, we present the focused analysis of German errors with the RStudio version 1.0.136 and the comparison with Spanish and English error words, if the analysis is available, for: length of the error words, error position, error categories and minimum number of edits.



Figure A.4: Distribution of the error categories for the German error resource [109].

We use the meta data “*misspelled length: number of characters the error word have*” [118] for the analysis of the German error length. The average length of phonemes for German ( $\sim 7.74$ ) and Spanish ( $\sim 7.47$ ) [126] error words is similar. Therefore, we design the auditory content with a focus on seven notes which intend to represent the seven phonemes.

We use the meta data “*error position: the position in the target word where the error occurs*” [118] for the analysis of the error position. Position 0 describes an error word with *multiple error positions*, and the *multiple error words* are segmented and separately annotated. Most German errors are at the beginning of a word which is the error position 1 ( $\sim 20.67\%$ ) or at error position 3 ( $\sim 21.06\%$ ) (see Figure A.3) while 70% of the Spanish errors are at the third position. Therefore, the third position is considered for the auditory attributes and the auditory content is designed with a focus on differences at the third position of, for example, a rhythm or series of notes.

We analysed the six error categories (*omission, substitution, multi-errors, insertion, boundary errors, and transposition*) from the

Damerau–Levenshtein distance	1	2	3	4	5
Total of 1021 error words	830	142	24	20	5
%	81.29	13.91	2.35	1.96	0.49

Table A.1: Distribution of the Damerau-Levenshtein distance [26, 78] for the German error words.

Spanish and English annotation of errors [126] and the two new error categories (*capital letter and non-capital letter errors*) from the analysis of German errors. For example, *substitution* is changing one letter for another [118], *omission* is missing a letter while *capital letter* describes a missing capital letter at the beginning of a word. *Omission* and *substitution* are the two most frequent error categories for German (*omission*,  $n = 290$ , 28.4%; *substitution*,  $n = 214$ , 20.96%, see Figure A.4), equal to Spanish and English [126]. Therefore, we focus on exchange music notes instead of letters (*substitution*) or leaving music notes out (*omission*) when designing the auditory content in Chapter 5.

We use the meta data “*Damerau–Levenshtein distance: the minimum number of edits (insertions, omissions, substitutions, transpositions) required to change the misspelled error into the (target) correct word [26, 78]*” [118] for the analysis of the error position. The distribution of the Damerau–Levenshtein distance for German errors shows with 81.3% a climax for one edits. Spanish error words have a similar high percentage for one letter mistakes (73.3%). The detailed results are given in Table A.1.

We made this additional small analysis of the German error words from [109, 118] to compare them with the summary of the Spanish and English annotations [126]. We aimed to identify obvious similar attributes between the error analysis and easy to deploy characteristics for the design of the auditory content in Chapter 5.

## A.3 Study Approvals

In the following, we show the approvals for our studies.

First, approval for *MusVis* was given by the Ministry of Education, Science and Culture of Schleswig-Holstein (*Ministerium für Bildung, Wissenschaft und Kultur, MBWK*, see Figure A.5) and by the *Education Authority* of Lower Saxony State (*Niedersächsische Landesschulbehörde*, see Figure A.6 and A.7).

Second, approval for *DGames* was given by the Ministry of Education, Science and Culture of Schleswig-Holstein (*Ministerium für Bildung, Wissenschaft und Kultur, MBWK*, see Figure A.8) and by the *Education Authority* of Lower Saxony State (*Niedersächsische Landesschulbehörde*, see Figure A.9 and A.10). The documents are required to be submitted and approved in German.

## A.4 Further Acknowledgements

Additionally to the acknowledgments at the beginning of this thesis, I would like to take the chance to thank significant individuals, groups, and supporters.

First, I would like to thank all teachers, students, and parents from the state of Lower Saxony for their participation and time! Special thanks goes to one class and one teacher which cannot be named due to the anonymous regulations, but they know who they are!

Vielen Dank an alle Lehrer/innen, Schüler/innen und Eltern aus Niedersachsen für eure Teilnahme. Aufgrund der Auflagen der Landesschulbehörde kann ich euch nicht persönlich nennen. Aber ihr wisst ihr seid gemeint. Besonders, möchte ich einer ganz besonderen Schule, Klasse und Lehrerin sowie der ansässigen Bibliotheksmitarbeiterin danken. Ihr wisst ich meine euch!!

I deeply thank for their support L. Albó, Barcelona; *ChangeDyslexia*, Barcelona; M. Jesús Blanque and R. Noé López,

school *Hijas de San José*, Zaragoza; A. Carrasco, E. Méndez and S. Tena, innovation team of school *Leonardo da Vinci*, Madrid; in Spain, and L. Niemeier, *Fröbel Bildung und Erziehung gemeinnützige GmbH*, Berlin; E. Prinz-Burghardt, *Lerntherapeutische Praxis*, Duderstadt; L. Klaus, *Peter-Ustinov-Schule*, Eckernförde; H. Marquardt, *Gorch-Fock-Schule*, Eckernförde; M. Batke and J. Thomaschewski, *Hochschule Emden/Leer*, Emden; N. Tegeler, *Montessori Bildungshaus Hannover gGmbH*, Hannover; Y. Schulz, *Grundschule Heidgraben*, Heidgraben; T. Westphal, *Leif-Eriksson-Gemeinschaftsschule*, Kiel; F. Goerke, *Grundschule Luetjensee*, Luetjensee; B. Wilke, *Schule am Draiberg*, Papenburg; P. Stümpel, *AncoraMentis*, Rheine; A. Wendt, *Grundschule Seth*, Seth; K. Usemann, *OGGS Meyerstraße*, Wuppertal; in Germany. Thanks to H. Witzel for his advice during the design of the visual part and to M. Blanca and M. Herrera for the Spanish version translation.

There are times when you search for collaborators or participants but your offers are turned down through one refusal after another. In times like these it is essential to have great supporters like Monika Batke, Friederike Hansch, Ute Lustig, Nancy Tegeler, Kirsten Usemann, Gisela Rauschenberger, Emilia Gómez, and Aurelio Ruiz García! Thank you so much for your kind words, your inspiration, and your support in spreading the word about my online experiments.

Thanks to the WSSC Group+ with Ana, Bora, Ch@to, Diana, Francesco, Marzieh, +Silvia, and Valerio, as well as previous members such as Çiğdem, Diego, Ioanna, and Lorena. Thanks for your company IP4EC+ with Adrian, Alexander, Antoine, Arash, Gabriela, Gino, Itziar, Praveen, Raquel, Syed, Trevor, and Xavier.

Without the people at UPF and especially the PhD students it would not be so much fun to work hard. During my time in Barcelona I spent a lot of time with all of you and your friendship meant a lot!

Finally, thanks to all parents and children for playing *DysMus*, *MusVis*, and *DGames*.

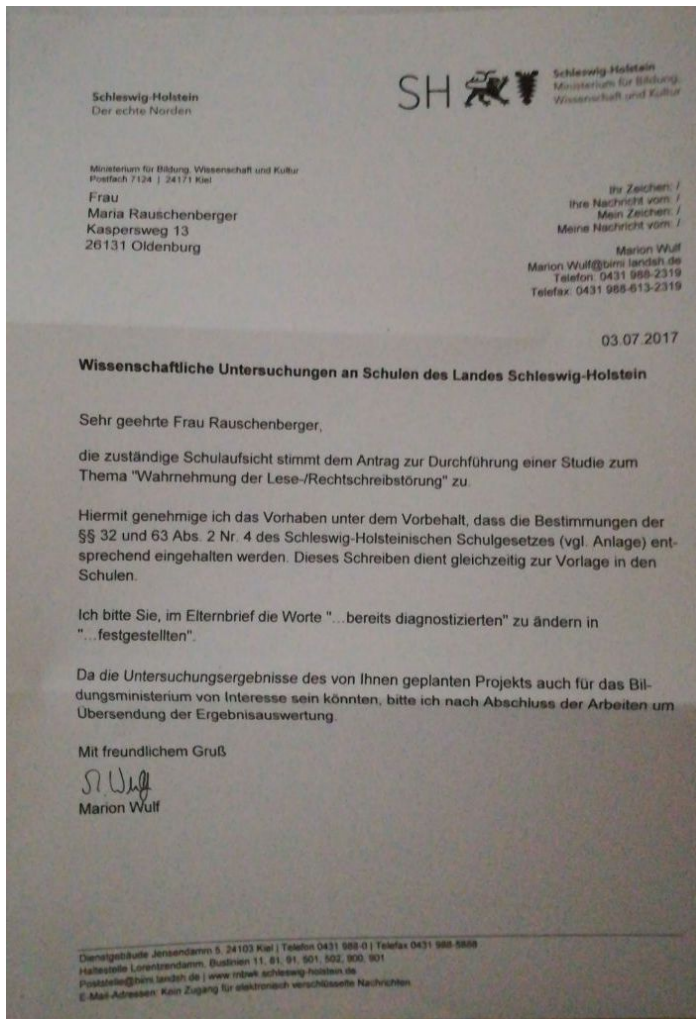


Figure A.5: MusVis study approval by the Ministry of Education, Science and Culture of Schleswig-Holstein (*Ministerium für Bildung, Wissenschaft und Kultur, MBWK*) in German.



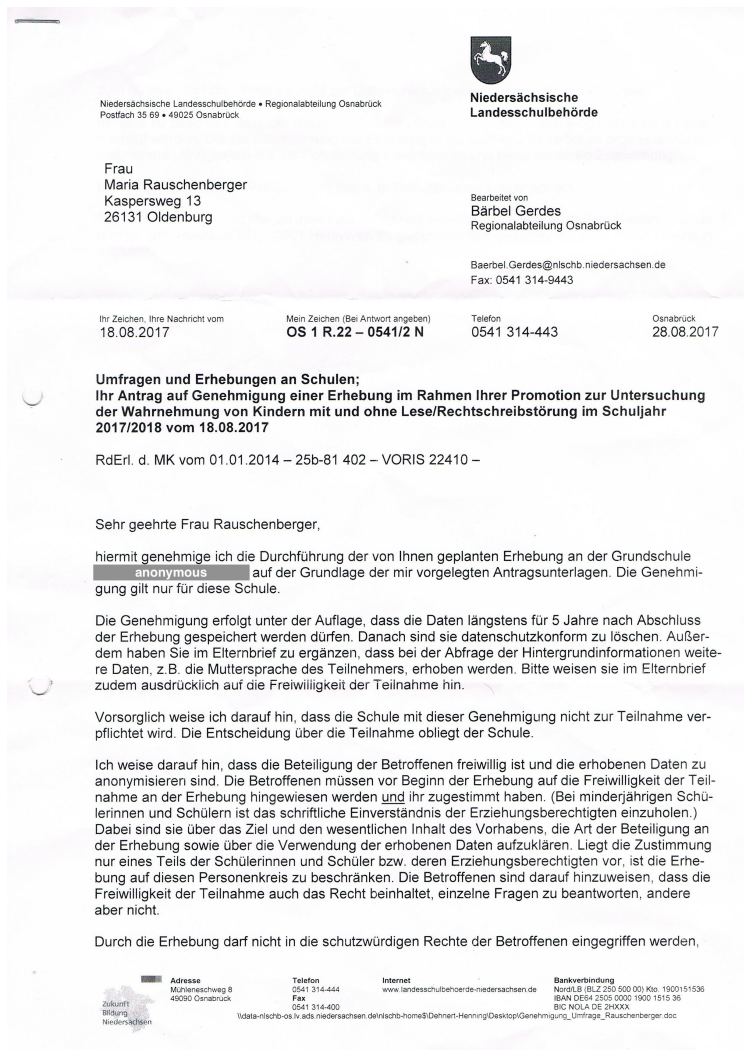


Figure A.6: MusVis study approval by the *Education Authority* of the State of Lower Saxony (*Niedersächsische Landesschulbehörde*) in German (part one).

- 2 -

zum Beispiel darf die Erhebung nicht zur Diskriminierung von einzelnen Personen führen.


Ich bitte zu beachten, dass der Unterricht und der übrige Schulbetrieb nach Möglichkeit nicht beeinträchtigt werden. Die zur Durchführung der Erhebung in der Schule erforderlichen organisatorischen Maßnahmen sind jeweils mit der Schulleitung abzustimmen und bedürfen deren Zustimmung.

Im Übrigen bitte ich die Ausführungen des o. g. Bezugserlasses zu beachten.

Für Ihre Erhebung wünsche ich Ihnen viel Erfolg und bitte Sie, mir sowie auch dem Nieders. Kultusministerium, Postfach 161, 30001 Hannover, zu gegebener Zeit das Ergebnis Ihrer Arbeit schriftlich mitzuteilen.

Mit freundlichen Grüßen

Im Auftrage



Dr. Henning Dehnert

Figure A.7: MusVis study approval by the *Education Authority* of the State of Lower Saxony (*Niedersächsische Landesschulbehörde*) in German (part two).

Schleswig-Holstein  
Der echte Norden



Schleswig-Holstein  
Ministerium für Bildung,  
Wissenschaft und Kultur

Ministerium für Bildung, Wissenschaft und Kultur  
Postfach 7124 | 24171 Kiel

Frau  
Maria Rauschenberger  
Kaspersweg 13  
26131 Oldenburg

Ihr Zeichen: /  
Ihre Nachricht vom: /  
Mein Zeichen: /  
Meine Nachricht vom: /

Marion Wulf  
Marion.Wulf@bimi.landsh.de  
Telefon: 0431 988-2319  
Telefax: 0431 988-613-2319

2. Februar 2018

### Wissenschaftliche Untersuchungen an Schulen des Landes Schleswig-Holstein

Sehr geehrte Frau Rauschenberger,

die zuständige Schulaufsicht stimmt dem Antrag zur Durchführung einer Studie zum Thema "Untersuchung der Wahrnehmung von Kindern mit und ohne eine Lese/Recht-schreibstörung" zu.

Hiermit genehmige ich das Vorhaben unter dem Vorbehalt, dass die Bestimmungen der §§ 32 und 63 Abs. 2 Nr. 4 des Schleswig-Holsteinischen Schulgesetzes (vgl. Anlage) entsprechend eingehalten werden. Dieses Schreiben dient gleichzeitig zur Vorlage in den Schulen.

Da die Untersuchungsergebnisse des von Ihnen geplanten Projekts auch für das Bildungsministerium von Interesse sein könnten, bitte ich nach Abschluss der Arbeiten um Übersendung der Ergebnisauswertung.

Mit freundlichem Gruß

  
Marion Wulf

Dienstgebäude Jensendamm 5, 24103 Kiel | Telefon 0431 988-0 | Telefax 0431 988-5888  
Haltestelle Lorentzendamm, Buslinien 11, 81, 91, 501, 502, 900, 901  
Poststelle@bimi.landsh.de | www.mbwk.schleswig-holstein.de  
E-Mail-Adressen: Kein Zugang für elektronisch verschlüsselte Nachrichten.

Figure A.8: DGames study approval by the Ministry of Education, Science and Culture of Schleswig-Holstein (*Ministerium für Bildung, Wissenschaft und Kultur, MBWK*) in German.



Frau  
Maria Rauschenberger  
Kaspersweg 13  
26131 Oldenburg

Bearbeitet von  
Bärbel Gerdes  
Regionalabteilung Osnabrück

Baerbel.Gerdes@nitschb.niedersachsen.de  
Fax: 0541 314-9443

Ihr Zeichen, Ihre Nachricht vom  
04.01.2018

Mein Zeichen (Bei Antwort angeben)  
OS 1 R.22 – 0541/2 N

Telefon  
0541 314-443

Osnabrück  
12.01.2018

**Umfragen und Erhebungen an Schulen;  
Antrag auf Genehmigung einer Erhebung im Rahmen Ihrer Promotion mit dem Titel:  
„Online-Studie zur Untersuchung der Wahrnehmung von Kindern mit und ohne eine Lese-  
/Rechtschreibstörung in Niedersachsen“**

RdErl. d. MK vom 01.01.2014 – 25b-81 402 – VORIS 22410 –

Sehr geehrte Frau Rauschenberger,

hiermit genehmige ich die Durchführung der von Ihnen geplanten Erhebung an der Grundschule  
**anonymous** auf der Grundlage der mir vorgelegten Antragsunterlagen. Die  
Genehmigung ist nur für diese Schule gültig.

Die Genehmigung erfolgt unter der Auflage, dass die Daten längstens für 5 Jahre nach Abschluss  
der Erhebung gespeichert werden dürfen. Danach sind sie datenschutzkonform zu löschen.

**Vorsorglich weise ich darauf hin, dass die Schule mit dieser Genehmigung nicht zur  
Teilnahme verpflichtet wird. Die Entscheidung über die Teilnahme obliegt der Schule.**

Ich weise darauf hin, dass die Beteiligung der Betroffenen freiwillig ist und die erhobenen Daten zu  
anonymisieren sind. Die Betroffenen müssen vor Beginn der Erhebung auf die Freiwilligkeit der  
Teilnahme an der Erhebung hingewiesen werden und ihr zugestimmt haben. (Bei minderjährigen  
Schülerinnen und Schülern ist das schriftliche Einverständnis der Erziehungsberechtigten  
einzuholen.) Dabei sind sie über das Ziel und den wesentlichen Inhalt des Vorhabens, die Art der  
Beteiligung an der Erhebung sowie über die Verwendung der erhobenen Daten aufzuklären. Liegt  
die Zustimmung nur eines Teils der Schülerinnen und Schüler bzw. deren Erziehungsberechtigten  
vor, ist die Erhebung auf diesen Personenkreis zu beschränken. Die Betroffenen sind darauf  
hinzuweisen, dass die Freiwilligkeit der Teilnahme auch das Recht beinhaltet, einzelne Fragen zu  
beantworten, andere aber nicht.

Durch die Erhebung darf nicht in die schutzwürdigen Rechte der Betroffenen eingegriffen werden,  
zum Beispiel darf die Erhebung nicht zur Diskriminierung von einzelnen Personen führen.

Ich bitte zu beachten, dass der Unterricht und der übrige Schulbetrieb nach Möglichkeit nicht



Adresse  
Mühlenschweg 8  
49090 Osnabrück

Telefon  
0541 314-444  
Fax  
0541 314-400

Internet  
www.landesschulbehoerde-niedersachsen.de

Bankverbindung  
Nord/LB (BLZ 250 500 00) Kto. 1900151536  
IBAN DE44 2505 0000 1900 1515 36  
BIC NOLA33HAN33  
R:\Umfragen\Sammlung\Rauschenberger\l.doc

Figure A.9: DGames study approval by the Education Authority of the State of Lower Saxony (*Niedersächsische Landesschulbehörde*) in German (part one).

- 2 -

beeinträchtigt werden. Die zur Durchführung der Erhebung in der Schule erforderlichen organisatorischen Maßnahmen sind jeweils mit der Schulleitung abzustimmen und bedürfen deren Zustimmung.

**Bei etwaigen Veröffentlichungen über dieses Vorhaben bitte ich sicherzustellen, dass Rückschlüsse auf die Schule, auf die Schulleitung, auf einzelne Lehrkräfte, auf einzelne Klassen und auf einzelne Schülerinnen und Schüler sowie deren Erziehungsberechtigte nicht möglich sind.**

Im Übrigen bitte ich die Ausführungen des o. g. Bezuserlasses zu beachten.

Für Ihre Erhebung wünsche ich Ihnen viel Erfolg und bitte Sie mir sowie auch dem Nieders. Kultusministerium, Postfach 161, 30001 Hannover, zu gegebener Zeit das Ergebnis Ihrer Arbeit schriftlich mitzuteilen.

Mit freundlichen Grüßen

Im Auftrage

Bärbel Gerdes

Figure A.10: DGames study approval by the *Education Authority* of the State of Lower Saxony (*Niedersächsische Landesschulbehörde*) in German (part two).