Screening Risk of Dyslexia through a Web-Game using Language-Independent Content and Machine Learning

Maria Rauschenberger
WSSC, DTIC
Universitat Pompeu Fabra &
Max Planck Institute for
Software Systems
rauschenberger@mpi-sws.org

Ricardo Baeza-Yates
Khoury College of
Computer Sciences
Northeastern University at SV
rbaeza@acm.org

Luz Rello
Dept. of Information Systems
and Technology
IE Business School
luz.rello@ie.edu

ABSTRACT

Children with dyslexia are often diagnosed after they fail school even if dyslexia is not related to general intelligence. In this work, we present an approach for universal screening of dyslexia using machine learning models with data gathered from a web-based language-independent game. We designed the game content taking into consideration the analysis of mistakes of people with dyslexia in different languages and other parameters related to dyslexia like auditory perception as well as visual perception. We did a user study with 313 children (116 with dyslexia) and train predictive machine learning models with the collected data. Our method yields an accuracy of 0.74 for German and 0.69 for Spanish as well as a F1-score of 0.75 for German and 0.75 for Spanish, using Random Forests and Extra Trees, respectively. To the best of our knowledge this is the first time that risk of dyslexia is screened using a language-independent content web-based game and machine-learning. Universal screening with language-independent content can be used for the screening of pre-readers who do not have any language skills, facilitating a potential early intervention.

CCS Concepts

•Human-centered computing → Field studies; User studies; *Empirical studies in accessibility;* **•Social and professional topics → People with disabilities; •Software and its engineering →** *Interactive games;*

Keywords

Dyslexia; Detection; Pre-Readers; Serious Games; Web-based Assessment; Universal Screening; Language-Independent; Visual; Auditory; Gamification.

1. INTRODUCTION

Dyslexia is a *specific learning disorder* which affects from 5% to 15% of the world population [1]. Children with dyslexia are often diagnosed with spelling and reading errors or after school failure, even if dyslexia is not related to general intelligence.

Generally, dyslexia manifestations can be observed when children reach a certain age and literary knowledge. Current approaches to screen (pre-)readers require expensive personnel, such as a professional therapist or special hardware such as fMRI scans [26]. Previous research have studied signs of dyslexia that are not related to reading and writing such as visual perception, short-term memory, executive functions or auditory perception [12]. These signs could be used to screen potential dyslexia in pre-readers and our work shows a possible approach for doing this by using machine learning with data coming from a language-independent content integrated in a web-based game. Our game has the potential of being easily accessible, making parents aware the potential risk of dyslexia to further look for more help, *e.g.*, a medical doctor or therapist.

The game and the user study is designed with the *human-centered design* framework [17] to collect the data set. This is relevant since collecting personal data is challenging because of privacy and trust issues [2, 7]. As a result, the final data sets are small and *small data* makes the prediction with machine learning models more difficult. That is, there is the risk of over-fitting or having a data set too small to be divided into meaningful test, training and validation sets.

We use standard machine learning classifiers like Random Forest with and without class weights, Extra Trees and Gradient Boosting from the *Scikit-learn* library for the prediction of dyslexia with small data sets.

Our main contributions are the user study results and the first web-based game for screening risk of dyslexia, based on language-independent content and using machine learning. To gather the data of this study, we had participants already diagnosed with dyslexia, instead of using pre-readers (younger children), since that would have required a long-term study. Our results show that the approach is feasible and that a higher prediction accuracy is obtained for German than for Spanish participants.

The rest of the paper is organized as follows: Section 2 covers the related work while Section 3 explains the rationale behind the game design. In Section 4 we cover the study methodology and in Sections 5 and 6 the predictive models and their results. We discuss the results in Section 7 finishing with conclusions and future work in Section 8.

W4A '20, April 20–21, 2020, Taipei, Taiwan

ISBN 978-1-4503-7056-1/20/04…$15.00

DOI: https://doi.org/10.1145/3371300.3383342

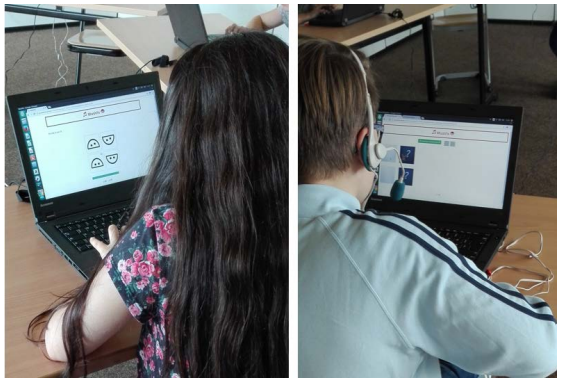

Figure 1: Participants playing the visual part (left) and the musical part (right) of MusVis. Photos included with the adults' permission.

2. RELATED WORK

Various applications and games to support, detect and treat dyslexia have been developed [31]. *Gamification* has been used to design various use cases, applications as well as frameworks [14, 22, 41, 47]. Gamification designs the *game play* of games with game elements to engage and motivate users [36, 42]. Games are developed to screen readers [39, 40] using linguistic content and to screen pre-readers [10, 11, 29, 30] focusing on the gameful experience. Only *Lexa* [27] published an accuracy (89.2%) using features related to phonological processing. However, they did not include game elements, and features are collected with extensive tests (extensive resources like time and cost). In addition, the classification is carried out on a small sample ($n = 56$), without any validation and no precautions or discussions about over-fitting.

Here we advance previous approaches by taking precautions on over-fitting, by not focusing on linguistic knowledge, and by using the same game content for every language. This will reduce the effort and time to design different content for different languages but more importantly, the content could be used for pre-readers.

3. GAME DESIGN

The aim of our web-game called *MusVis* (see Figure 1) is to measure the reaction of children with and without dyslexia while playing, in order to find differences on their behavior. A video of MusVis is available at http://bit.ly/MusVisContent. We designed our game with the assumption that non-linguistic content like rhythm or frequency [27] can represent the difficulties that a child with dyslexia has with writing and reading [12, 53], and dyslexia can be measured through the interaction of a person [39, 40] like total number of clicks or play duration.

An early game design and content was previously tested with a five-user study [23] to eliminate major usability mistakes which could have an influence on the prediction and validated our game measurements [34]. In our pilot study ($n = 178$), we found that there were four significant game measurements for Spanish, German, and English as well as eight significant game measurements for Spanish [32].

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

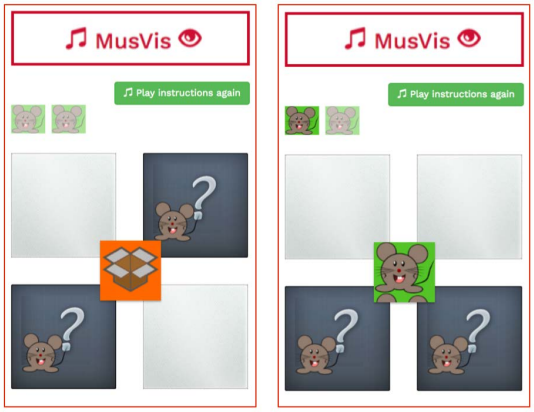

Figure 2: Example of the auditory part from the game *MusVis* for the first two clicks on two sound cards (left) and then when a pair of equal sounds is found (right). The participant is asked to find two equal auditory cues by clicking on sound cards.

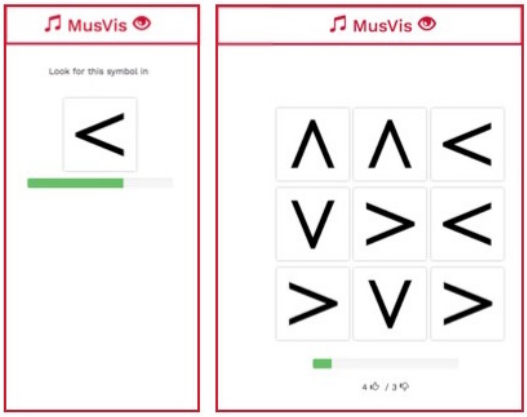

Figure 3: Example of the visual part of the game *MusVis* with the priming of the target cue *symbol* (left) and the nine-squared design including the distractors for each *symbol* (right).

We designed the language-independent game content taking into account the knowledge of previous literature selecting the most challenging content for people with dyslexia. Therefore, we designed language-independent content with auditory and visual cues.

We designed an auditory part (see Figure 2) and a visual part (see Figure 3) using features extracted from the literature. The game play is different due to the unequal perception of auditory and visual cues but each part targets in both cases general skills, *e.g.*, short-term memory [25].

As is well known, children have more difficulty paying attention over a longer period of time. Therefore, the two parts have four stages which are counter-balanced with *Latin Squares* [9]. Each stage has two rounds, which sums up to 16 rounds in total for the whole game. Each stage first has a round with four cards and then with six cards, needing less

Key	Name	Description
CS	Complex *vs.* simple	Children with dyslexia (DG) recall significantly fewer items correctly in a lab study for long memory spans [12]. The rhythmic complexity did not have an effect on the difference between DG and children without dyslexia (CG) [16].
Pi	Pitch	Pitch perception is essential for prosodic performance [16], is correlated to language development, and can be used as a predictor for language [53].
SD	Sound duration	Acoustic parameter differences in short tones (< 350 ms) are difficult to distinguish for a person with language difficulties [25].
RT	Rise time	Both groups showed significant differences when comparing *rise time* [12]. Rise time and prosodic development are strongly connected and were shown to be most sensitive to dyslexia [16].
Rh	Rhythm	DG show deficits in recalling the patterns of auditory cues [25]. However, rhythm modulations show no effect on the children performance [16].
STM	Short-term memory	DG show weaknesses in short-term memory tasks [25] when more items are presented [12]. Also, deficits can be frequently observed for the short-term auditory memory span [19].
PSE	Phonological similarity effect	DG have difficulties with similar sounds and the *phonological neighborhood* when long memory spans are used [12].
CAPS	Correlated acoustic parameters speech	Since the *phonological grammar* of music is similar to the prosodic structure of language, music (*i.e.,* a combination of acoustical parameters) can be used to imitate these features [53]. DG are "*reliably impaired in prosodic tasks*" [12].

Table 1: Description of the auditory attributes which show promising relations to the prediction of dyslexia.

Attributes	Auditory					General		
	CS	**Pi**	**SD**	**RT**	**Rh**	**STM**	**PSE**	**CAPS**
Literature								
Goswami et al.[12]	✓			✓		✓	✓	✓
Huss et al.[16]	✓	✓		✓	✓			
Johnson [19]						✓		
Overy [25]			✓		✓	✓		
Yuskaitis et al. [53]		✓						✓
Stage								
Frequency	✓	✓	✓			✓	✓	✓
Length			✓			✓	✓	✓
Rise time	✓		✓	✓		✓	✓	✓
Rhythm	✓		✓		✓	✓	✓	✓

Table 2: 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*.

than 10 minutes to play. We aim to address participants' motivation for both game parts with the design of the following game mechanics frequently used in learning environments [36]: rewards (points), feedback (instant feedback) or challenges (time limit), plus the game components (story for the game design).

The content design, user interface, interaction and implementation for the auditory and visual parts of the game are described in the following sections.

3.1 Auditory Game Design

The auditory part is inspired in the traditional game *Memory* in which pairs of identical cards (face down) must be identified by flipping them over [52]. 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, we used acoustic parameters; for example, to imitate the *prosodic* structure of language which is similar to the *phonological grammar* of music [28].

Musicians with dyslexia score better on auditory perception tests than the general population, but not on auditory working memory tests [21]. 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 [16] while taking the precaution of not measuring hearing ability [8]. Each stage is assigned to one acoustic parameter like frequency or rhythm which is designed with the knowledge of the analysis from previous literature [32].

Therefore, we used the acoustic parameters *frequency, length, rise time and rhythm* as auditory cues. Each auditory cue was assigned to a game stage (see Table 2), which we mapped to the attributes and literature references (see Table 1) that provide evidence for distinguishing a person with dyslexia.

For example, our *rhythm* stage uses the following characteristics: *complex vs. simple [12, 16], sound duration, rhythm [16], short-term memory [12, 19], phonological similarity effect [12], and correlated acoustic parameters speech [12, 53].*

Each acoustic stage has three auditory cues (we use MP3 for sound files). 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).

The auditory cues are generated with a simple sinus tone using the free software *Audacity*.[1] The exact parameters of each auditory cue are already published [32] and the auditory cues are available at *GitHub*. [2] [33] Each stage has two rounds, with first two and then three auditory cues that must be assigned by choosing the same sound (see Figure 2). The arrangement of sounds (which auditory cue matches which card) is random for each round.

3.2 Visual Game Design

The visual game play uses a Whac-A-Mole interaction similar to the first round of *Dytective* [39]. 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/.[3] We adapted the interaction design and content for this purpose (see Figure 3). For the visual game, we designed cues that have 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 [35, 38, 51].

To create the visual cues, we designed different visual

[1] *Audacity* is available at http://audacity.es/, Last access: May 2019.

[2] https://github.com/Rauschii/DysMusicMusicalElements

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

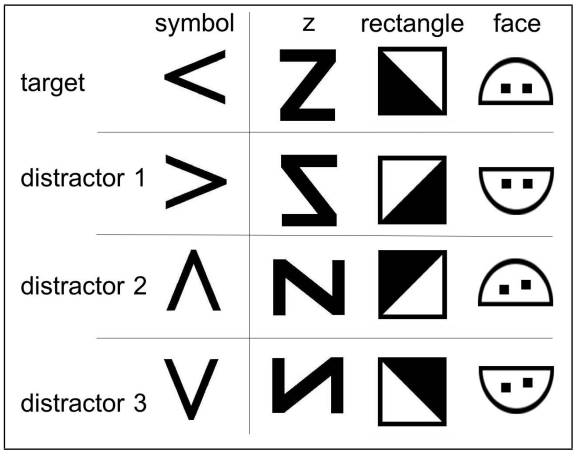

Figure 4: Overview of the designed visual cues. The figure shows the target cue (top) and distractor cues (below) for the four different stages (*z, symbol, rectangle, face*) of the visual part of the game *MusVis*.

representations similar to visual features of annotated error words from people with dyslexia [35, 38, 51] and designed the game as a simple search task, which does not require language acquisition.

In the beginning, participants are shown the target visual cues (see Figure 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 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* [9]. 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, top). The other three visual cues are *distractors* for the participants. Each stage has two rounds with first a 4-squared and then a 9-squared design (see Figure 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.

4. USER STUDY METHODOLOGY

We use the human-centered design framework to design our study and to collect the data for the prediction of dyslexia. We conducted a within-subject design study ($n = 313$) which means that all participants played all game rounds [9] with the same language-independent content. Only the game instructions were translated into each native language.

Spanish participants diagnosed with dyslexia were mainly recruited from public social media calls by non-profit organizations. 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 is international. The control groups for Spanish and German were recruited mostly with the collaboration of four schools, two in each country.

4.1 Online Data Collection

Collecting data is costly in terms of time consumption and privacy issues, specially if the data is related to education and health. Therefore, we must make the best of the limited resource [2, 7]. In our case, we need a certain age range to make sure a person with dyslexia is already diagnosed and has not been fully treated yet. Since our collected data is considered *small data* [2, 7], we need to analyze them accordingly, *i.e.*, avoid over-fitting using cross-validation instead of training, test and validation sets as well as using classifiers configured to avoid over-fitting.

4.2 Procedure and Ethics Statement

First, the parents were informed about the purpose of the voluntary study. Next, only after the parents gave the consent, children were allowed to participate in this user study from home or from school, with the first author of this work present or always available through digital communication.

The data collection for this user study has been approved by the German Ministry of Education, Science and Culture in Schleswig-Holstein (*Ministerium für Bildung, Wissenschaft und Kultur)* and Lower Saxony State Education Authority (*Niedersächsische Landesschulbehörde*). In Spain governmental approval was not needed in addition to the school approval.

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

After that participants played both parts of the game. At the end, two feedback questions are 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 participants data, if given. On the other hand, the name of the child is not collected and all data is stored on a password secured web server.

4.3 Participants

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 predictive models, we took 313 participants into account, including the 178 participants from the pilot study [32]. To have precise data, we took out participants that reported in the background questionnaire that they suspected of having dyslexia but did not have a diagnosis ($n = 48$).

The remaining participants were classified as diagnosed with dyslexia (DG) or not showing any signs of dyslexia (control group, CG), 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 Ger-

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

Table 3: Overview of the participants per data set.

man participants (DE, $n = 149$), and one for all languages (ALL, $n = 313$) in which we included participants that spoke English ($n = 11$). Participants ranged in age from 7 to 12 years old. The users in the data sets are described in Table 3. 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 the non-profit organization *ChangeDyslexia*. [4] 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 [20]. 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.

4.4 Dependent Variables and Features

The participant features are detailed in Table 4 while the dependent variables collected through the game are listed in Table 5. These variables were used for the statistical comparison of the pilot study and for the selection of the features for the predictive models. Feature 3 was set with the language selected for the instructions. Features 1, 2, 4 to 8 were answered with the online questions by the participants' supervisor. Feature 9 was collected from the browser during the study experiment.

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.

[4]https://changedyslexia.org/

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

Table 4: Description of participant features.

- *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.

Both parts

- *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.

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

Table 5: 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.

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 (see Table 5).

5. PREDICTIVE MODELS SETUP

In this section we present the machine learning techniques used 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 feature selection.

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), which are described in the Scikit-learn version 0.21.2 [46]. We address the risk of over-fitting on 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 [46]. While we have small data, we are not optimizing the input parameters of classifiers until we can hold out a test data set as proposed by scikit-learn 0.21.2 documentation to evaluate the changes [45] and to avoid biases [50]. To explore the best prediction conditions we used the feature selection as described in the next section.

5.2 Informative Features

We address the danger of selecting the correct features [18] by taking into account the knowledge of previous literature about the differences of children with an without dyslexia. For example, since there are two theories of the cause of dyslexia (visual *vs.* auditory [6]), we use subsets of visual and auditory features to explore the influence on the classifiers.

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 clicks, 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).

6. RESULTS

We followed the same steps of the pilot study to compare the statistical findings before giving the machine learning results.

6.1 Statistical Validation

The pilot study collected data from 178 participants (which were later included into our current data set, $n = 313$) to find significant differences on the game measurements [32]. Therefore, we apply first the *Shapiro-Wilk Test* and then the *Wilcoxon Test* since all game measures are not normally distributed. We use the Bonferroni correction ($p < 0.002$) to avoid type I errors. We present the results of the statistical analysis for the validation data ($n = 313$) separated by language and for all languages (see Table 6.1). Additionally, we compare the statistical analysis results from the pilot-study ($n = 178$) with the new 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 ($n = 178$). 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 seven significant variables that distinguish a person with or without dyslexia.

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 ($n = 149$).

Part	Data set	Variable	Control		Dyslexia		Mann-Whitney U		
			mean	sd	mean	d	W	p-value	effect size
Visual	ALL	**Total clicks**	6.8	2.7	7.2	3.2	670194	**2e-04**	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	**3e-08**	**0.31**
		First click	2.63s	1.69s	2.26s	1.22s	141938	**1e-04**	0.27
		Hits	5.8	3	6.5	2.9	136904	**2e-06**	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	**1e-04**	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
Auditory	ES	Total clicks	11.3	6	10.9	5.5	157282	0.15	0.07
		4th click	2.0s	1.3s	1.7s	1.0s	131228	**1e-08**	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	**2e-04**	0.19
		Average	3.0s	2.7s	2.6s	0.9s	121966	**5e-13**	0.29
	DE	Total clicks	11.1	5.5	11.5	6.6	166340	0.27	0.07
		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 6: Overview of dependent variables for visual (top) and auditory (below) features of *MusVis*. Significant results are in bold.

Model	Data	Feat.	Recall	Precis.	F1	Acc.
RF	**DE**	**5**	0.77	0.78	**0.75**	**0.74**
RFW	DE	5	0.75	0.75	0.74	0.73
Baseline	DE		0.60	0.37	0.46	0.50
ETC	**ES**	**20**	0.76	0.76	**0.75**	**0.69**
RF	ES	5	0.74	0.73	0.72	0.65
Baseline	ES		0.68	0.46	0.55	0.50
GB	**ALL**	**20**	0.66	0.65	**0.65**	**0.61**
GB	ALL	5	0.64	0.64	0.63	0.59
Baseline	ALL		0.63	0.40	0.49	0.50

Table 7: Best results of the different classifiers, features and data sets. Results are ordered by the best F1-score and accuracy.

6.2 Predictive Results

We processed our data sets with different classifiers and different subsets of features, following the description from the previous section.

We computed the *balanced accuracy* for our binary classification problem to deal with imbalanced data sets; for example, the ALL data set has dyslexia 37% *vs.* control 63%. The Dummy Classifier is computed for our imbalanced data with the most frequent label and reported with the balanced accuracy [46]. We do not apply over- or under-sampling to address our imbalanced data because the variances among people with dyslexia are broad, for example, difficulty level or the individual causes for perception differences.

As described in the previous section 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 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 *balanced accuracy* scores for each data set as well as the baseline, as can be seen in Table 7. We outperform our baseline 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 0.1 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 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 0.1 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 normalized confusion matrix (see Figure 5) does not show over-fitting for the best results for DE, ES and ALL. The fact that the best results are with few features imply that the rest are highly correlated or noisy.

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 6. For example, reducing the features for DE improves the accuracy for ET,

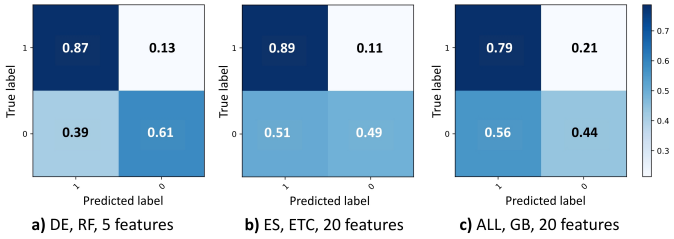

Figure 5: Normalized confusion matrix for the three best results (F1-score and accuracy): a) *DE, 5 features with RF*; b) *ES, 20 features with ETC*; and c) *ALL, 20 features with GB.*

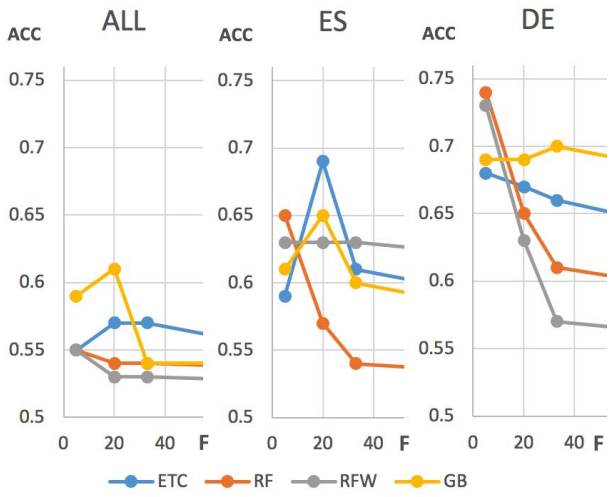

Figure 6: The plot shows the relation of accuracy to features for all classifiers in the data set ALL (left), ES (middle) and DE (right).

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.

7. DISCUSSION

Most children with dyslexia show a varying severity of deficits in more than one area [4], which makes dyslexia more a spectrum than a binary disorder. Additionally, we rely on current diagnostic tools (*e.g.,* DRT [13, 48]) 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 that dyslexia has.

7.1 Group Comparison

The measurement data taken from the game *MusVis* show that Spanish participants with dyslexia behave differently than their control group. Differences can be reported for the auditory game part for: *4th 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.*

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 had eight significant indicators in the pilot study 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 [1]. The assumption that the tendencies for the indicators are similar over all languages cannot (yet) 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 one limitation: participants could select a correct pair by chance, *e.g.,* 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** [5, 44]. Therefore, we designed our game with content that is known to be difficult 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 [39]. 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 *game play.*

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** [49] or recall the auditory and visual information from short-term memory [12]. 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** (*e.g.*, more musical training).

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** [12] 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 medium effect sizes of 0.29, so we can estimate what the effects would be in the whole population [9].

A person with dyslexia has difficulties with reading and writing independent of the mother tongue, which also appear when learning a second language [15, 24]. The analysis of errors from children with dyslexia show similar error categories for Spanish, English [38], and German [35], revealing similarities of perception between the languages.

Our results from the pilot study [32] suggest that we can measure a significant difference on four indicators for the visual game with the same tendency between Spanish, German, and English. With all our data ($n = 313$), we can confirm just one significant dependent variable with the same tendency for Spanish and German.

Still 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.

7.2 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 accuracy 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 for pre-readers.

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. However to avoid over-fitting we did not modify the default value for the threshold (typically 0.5), something that we plan to study in the near future as we need to increase recall for the dyslexia class keeping a reasonable number of false positives.

If a person with dyslexia is not discovered (early), they are prone to face additional issues such as anxiety, sadness and decreased attention [43]. 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, both, as a preventive measure and as 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 [6]. 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 in different people. We might be able to measure stronger effects when we design visual and auditory cues that have more attributes related to dyslexia, including some that favor the latter.

The ALL data set reached *only* an accuracy of 0.61, which might be due to the following reasons. 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 by bilingualism. It is argued that a person who speaks more than one language has more knowledge of their first language than a monolingual person [20], 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 [37]). 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, taking into account 27 features (GB) performs no better than using 20 features (GB) for 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 to understand the data and results better. In our case, we see that ALL does not perform 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 [3], something that made sense in retrospect.

8. CONCLUSIONS AND FUTURE WORK

We processed our game data with Extra Trees, Random Forest without and with class weights, and Gradient Boost to predict dyslexia using a data set of 313 participants. We reached the best accuracy of 74% for the German case using RF while the best accuracy for Spanish was 69% using ETC.

Our approach can optimize resources for detecting and treating dyslexia, however, it would need at the beginning more personnel to screen many more children at a young age to enlarge our training data. As children with dyslexia need around two years to compensate their difficulties, our approach could help to decrease school failure, late treatment and most importantly, to reduce suffering for children and parents.

The main advantage of our language-independent content approach is that has the potential to screen pre-readers in the near future. Indeed, we aim to collect more data with younger children to improve our results.

Future work includes improving our machine learning models and do further feature analysis. More explainable models should also be considered.

9. ACKNOWLEDGMENTS

This paper and content were partially funded 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*. First, we 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. We 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.

We also thank all parents and children for playing *MusVis*. Finally, thanks to H. Witzel for his advice durithe development of the visual part and to M. Blanca, and M. Herrera for the translation of the Spanish version.

References

[1] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, London, England, May 2013.

[2] 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.

[3] A. Bandhyopadhyay, D. Dey, and R. K. Pal. *Prediction of Dyslexia using Machine Learning — A Research Travelogue*, volume 24. Springer Singapore, 2018.

[4] 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.

[5] 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.

[6] G. De Zubicaray and N. O. Schiller. *The Oxford handbook of neurolinguistics*. Oxford University Press, New York, NY, 2018.

[7] J. J. Faraway and N. H. Augustin. When small data beats big data. *Statistics & Probability Letters*, 136:142–145, May 2018.

[8] H. Fastl and E. Zwicker. *Psychoacoustics*. Springer Berlin Heidelberg, Berlin, Heidelberg, third edition, 2007.

[9] Field et al. *How to design and report experiments*. SAGE Publications, London, 2003.

[10] 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):1–24, Apr 2017.

[11] 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, Switzerland, 2015.

[12] 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.

[13] 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.

[14] 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.

[15] T. Helland and R. Kaasa. Dyslexia in English as a second language. *Dyslexia*, 11(1):41–60, Feb 2005.

[16] 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.

[17] 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.

[18] 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.

[19] D. J. Johnson. Persistent auditory disorders in young dyslexic adults. *Bulletin of the Orton Society*, 30(1):268–276, Jan 1980.

[20] I. Kecskes and T. Papp. *Foreign Language and Mother Tongue*. Psychology Press, New York, 1 edition, Jun 2000.

[21] 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.

[22] A. Mora, D. Riera, C. Gonzalez, and J. Arnedo-Moreno. A Literature Review of Gamification Design Frameworks. In *7th International Conference on Games and Virtual Worlds for Serious Applications*, 2015.

[23] J. Nielsen. Why You Only Need to Test with 5 Users. *Jakob Nielsens Alertbox*, 19(September 23):1–4, 2000.

[24] J. Nijakowska. *Dyslexia in the foreign language classroom*. Multilingual Matters, 2010.

[25] 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.

[26] 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.

[27] 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.

[28] R. F. Port. Meter and speech. *Journal of Phonetics*, 31:599–611, 2003.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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, June 2017.

[34] 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.

[35] 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).

[36] 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.

[37] 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.

[38] 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.

[39] L. Rello, M. Ballesteros, A. Ali, M. Serra, D. Alarcón, and J. P. Bigham. Dytective: Diagnosing Risk of Dyslexia with a Game. In *Pervasive Health 2016*, pages 89–96, Cancun, Mexico, May 2016. ACM Press.

[40] 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.

[41] 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.

[42] R. Rouse. *Game Design: Theory and Practice, Second Edition: Theory and Practice, Second Edition*. Wordware Publishing, Inc., 2004.

[43] G. Schulte-Körne. Diagnostik und Therapie der Lese-Rechtscheib-Störung (The prevention, diagnosis, and treatment of dyslexia). *Deutsches Ärzteblatt international*, 107(41):718–727, 2010.

[44] 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.

[45] Scikit-learn. 3.1. Cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html, 2019.

[46] Scikit-learn Developers. Scikit-learn Documentation. https://scikit-learn.org/stable/documentation.html.

[47] K. Seaborn and D. I. Fels. Gamification in theory and action: A survey. *International Journal of Human Computer Studies*, 74:14–31, 2015.

[48] C. Steinbrink and T. Lachmann. *Lese-Rechtschreibstörung (Dyslexia)*. Springer Berlin Heidelberg, 2014.

[49] P. Tallal. Improving language and literacy is a matter of time. *Nature reviews. Neuroscience*, 5(9):721–728, 2004.

[50] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, Feb 2006.

[51] 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.

[52] Wikipedia. Memory (Spiel) (Memory Game), 2019.

[53] 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.