Measuring reading ability in the web-browser with a lexical decision task

Jason D. Yeatman*[1,2], Kenny An Tang*[1,2], Patrick M. Donnelly[3,4], Maya Yablonski[1,2,5], Maha Ramamurthy[1,2], Iliana I. Karipidis[6], Sendy Caffarra[1,2,7], Megumi E. Takada[1,2], Michal Ben-Shachar[5,8], Benjamin W. Domingue[2]

1) Stanford University School of Medicine, Division of Developmental-Behavioral Pediatrics, Stanford, CA, USA
2) Stanford University Graduate School of Education, Stanford, CA, USA
3) University of Washington, Institute for Learning & Brain Science, Seattle, WA, USA
4) University of Washington, Department of Speech & Hearing Sciences, Seattle, WA, USA
5) The Gonda Multidisciplinary Brain Research Center, Bar Ilan University, Ramat-Gan, Israel
6) Center for Interdisciplinary Brain Sciences Research, Department of Psychiatry and Behavioral Sciences, School of Medicine, Stanford University, Stanford, CA, USA
7) Basque Center on Cognition, Brain and Language, San Sebastian, Spain
8) Department of English Literature and Linguistics, Bar-Ilan University, Ramat-Gan, Israel

*These authors contributed equally to this work
Corresponding author: Jason D. Yeatman, Ph.D. - jyeatman@stanford.edu

Abstract

An accurate model of the factors that contribute to individual differences in reading ability depends on data collection in large, diverse and representative samples of research participants. However, that is rarely feasible due to the constraints imposed by standardized measures of reading ability which require test administration by trained clinicians or researchers. Here we explore whether a simple, two-alternative forced choice, time limited lexical decision task (LDT), self-delivered through the web-browser, can serve as an accurate and reliable measure of reading ability. We found that performance on the LDT is highly correlated with scores on standardized measures of reading ability such as the Woodcock-Johnson Letter Word Identification test administered in the lab ($r = 0.91$, disattenuated $r = 0.94$) . Importantly, the LDT reading ability measure is highly reliable ($r = 0.97$). After optimizing the list of words and pseudowords based on item response theory, we found that a short experiment with 80 words (2-3 minutes) provides a reliable ($r = 0.95$) measure of reading ability. Thus, the self-administered, rapid online assessment of reading ability (ROAR) developed here overcomes the constraints of resource-intensive, in-person reading assessment, and provides an efficient and automated tool for effective online research into the mechanisms of reading (dis)ability.

Why is learning to read almost effortless for some children and a continuous struggle for others? Understanding the mechanisms that underlie individual differences in reading abilities, and the factors that contribute to reading disabilities, is an important scientific challenge with far reaching practical implications. There is near consensus on the importance of foundational skills like phonological awareness for reading development [1–3]. However, there are a myriad of other sensory, cognitive and linguistic abilities whose link to reading is fiercely debated [4–11]. For example, hundreds of papers, published over the course of four decades, have debated the causal link between deficits in the magnocellular visual pathway and developmental dyslexia [6,12–18]. Similarly contentious is the link between rapid auditory processing and developmental dyslexia [19–23]. More recently, the role of various general learning mechanisms (e.g., statistical learning, sensory adaptation) in reading development has become a topic of great interest [24–28].

One reason these debates persist is that most findings reflect data from relatively small samples collected in a single laboratory and, therefore, findings are not necessarily representative of the population [29–31]. There are two major factors that limit the feasibility of collecting data in large, diverse, representative samples. First, experiments measuring specific sensory and cognitive processes typically depend on experimental software installed on computers in a laboratory. Second, standardized measures of reading ability typically require a trained test administrator to administer the test to each research participant. These two factors: (a) make it time consuming and costly to recruit and test large samples, (b) prohibit including research subjects that are more than a short drive from a university, (c) bias samples towards university communities and (d) create major barriers for the inclusion of under-represented groups.

Over the past decade, advances in software packages for running experiments through the web-browser have catalyzed many areas of psychological research to pursue larger, more diverse and representative samples. For example, software packages such as jsPsych [32,33], Psychopy [34], Gorilla [35] and others, make it feasible to present stimuli and record behavioral responses in the web-browser with temporal precision that rivals many popular software platforms for collecting data in the laboratory [36]. However, online experiments have yet to make major headway in the developmental and educational spheres, specifically in reading and dyslexia research. Standardized testing of reading ability still depends on individually administered tests. Thus, even though a researcher might be able to accurately measure processing speed or visual motion perception in thousands of subjects through the web-browser, they would still need to individually administer standardized reading assessments to each participant.

Our goal here was to develop an accurate, reliable, expedient and automated measure of single word reading ability that could be delivered through the web-browser. Specifically, we sought to design an experiment to approximate scores on the Woodcock-Johnson Letter Word Identification (WJ-Word-ID) test [37]. The WJ-Word-ID is one of the most widely used standardized measures of reading ability in research (and practice). Like other test batteries that contain tests of reading ability (e.g., The NIH Toolbox [38], Wide Range Achievement Test [39], Wechsler Individual Achievement Test [40]), the WJ-Word-ID requires that an experimenter or clinician present each research participant with single printed words for them to read out loud. Participant responses are manually scored by the test administrator as correct or incorrect based

on accepted rules of pronunciation. The words are organized in increasing difficulty such that the number of words that a participant reads correctly is also a measure of the difficulty of words that the participant can read. Due to excellent psychometric properties, and widely accepted validity, the Woodcock-Johnson and a compendium of similar tests are at the foundation of thousands of scientific studies of reading ability.

In considering the myriad of tasks that might produce a suitable browser-based measure of reading ability, the lexical decision task (LDT) is a good candidate for both theoretical and practical reasons. The LDT has a rich history in the cognitive science literature as means to probe the cognitive processes underlying visual word recognition. It is broadly assumed that many of the same underlying cognitive processes are at play when participants make a decision during a two-alternative forced choice (2AFC) lexical decision as during other word recognition tasks (e.g. naming [41,42]). Unlike naming or reading aloud, the LDT can be: (a) scored automatically without reliance on (still imperfect) speech recognition algorithms, (b) completed in group (or public) settings such as a classroom and (c) administered quickly as each response only takes, at most, 1-2 seconds.

There is empirical support for LDT performance being linked to reading ability. For example, Martens and de Jong demonstrated that response times (RT) on the LDT differed between children with dyslexia compared to children with typical reading skills [43]. Specifically, they found larger word-length and lexicality effects in children with dyslexia compared to children with typical reading skills. These effects seem to track reading ability, as opposed to being a specific marker of dyslexia: reading matched controls, i.e., younger children matched to the children with dyslexia in terms of raw reading scores, showed similar word-length and lexicality effects as found in the older children with dyslexia. Similar to our goal here, Katz and colleagues examined the LDT in combination with a Naming Task, as predictors of standardized reading measures [44]. They found that average response times on the LDT (when combined with the Naming Task) correlated with a variety of reading ability composite scores, including the Woodcock Johnson Basic Reading Skills [44]. However, their study primarily looked at young adults (median age of 21.5 years), and focused on RT as the independent variable, since most participants were at ceiling in terms of accuracy. It is not clear how well these effects would generalize to children learning to read. Moreover, effect sizes were relatively small, suggesting that RT on the LDT may not be the best predictor of reading ability, or that the range of stimuli was not wide enough. Here we sought to develop an implementation of the LDT in the web-browser, along with a suitable list of words and pseudowords, to efficiently and accurately measure reading ability in a broader age group, spanning first grade through young adulthood.

The present work (1) reports the results from a first study in which we developed, validated and optimized a browser-based LDT-based measure of reading ability (*Study 1*) and (2) defines the plan for a second larger-scale norming study (*Study 2*). *Study 1* was an exploratory study in which we employed a long list of words and pseudowords, spanning a broad range of lexical and orthographic properties. The goal of *Study 1* was to, first, assess the feasibility of using a browser-based LDT to estimate reading ability and, second, leverage the item level data in combination with item response theory (IRT) to construct a collection of short-form tests with optimal psychometric properties. *Study 2* is planned to collect normative

data in a large and representative sample of elementary school children based on the optimized test from *Study 1*. Based on the larger sample in *Study 2*, we will more accurately measure the relationship between LDT task-performance and various standardized measures of reading ability that are commonly used in research (and practice).

STUDY 1: METHODS

Browser-based lexical decision task (version 1)

A visual-presentation 2AFC lexical decision task (LDT) was implemented in the Psychopy experiment builder [34], converted to Javascript, and uploaded to Pavlovia, an online experiment platform [45].

The LDT was split into five blocks, each consisting of 100 trials (50 real words, 50 pseudowords) and each block was introduced as a quest in the magical world of Lexicality. To engage a wide range of subjects, the LDT was embedded in a fun story, animated environment, and a game that involved collecting golden coins. Prior to the five blocks of the LDT, subjects completed a simple response time (RT) task. This task was used as a baseline measure, in case there was too much noise in the response time data (e.g., due to different devices being used). In each trial of the simple RT task, participants saw a fixation cross, then a triangle or a square flashed briefly for 300 ms, followed by another fixation cross. Participants were instructed to use the keyboard to respond, by pressing "left arrow" to indicate that a triangle was presented, or "right arrow" to indicate that a square was presented.

The five blocks of the LDT followed a similar format: a word or pseudoword would flash briefly for 300 ms in the center of the screen, followed by a fixation cross. Participants pressed "right" for a real word or "left" for a pseudoword. Stimuli were presented in the center of the screen and scaled to be 3.5% of the participants' vertical screen dimension.

The task was gamified in order to be more engaging for the younger participants. If the participants achieved ten consecutive correct answers, a small, one-second animation would play showing their party of adventurers defeating a group of monsters, with a small "+10 gold" icon appearing on top. After each block, participants unlocked new characters and were shown the amount of gold that they had collected during the last block.

Word Stimuli

250 real words and 250 pseudowords were selected to span a large range of lexical and orthographic properties, with the goal of finding stimuli spanning a large range of difficulty. The distribution of lexical and orthographic properties of the stimuli is shown in **Figure 1**. Pseudowords were matched in orthographic properties to real words using Wuggy, a pseudoword generator [46]. Wuggy generates orthographically plausible pseudowords that are matched in terms of sub-syllabic segments, word length, and letter-transition frequencies. Matched real word/pseudoword pairs were kept within the same block, such that each block contained 50 real words and 50 pseudowords roughly matched in terms of orthographic

properties. Additionally, 18 real words were paired with 18 orthographically implausible pseudowords to create some easier items to ensure that the LDT would not have floor effects for young children and those with dyslexia (e.g. "wndo"). These orthographically implausible words had low trigram frequencies and can be appreciated from the bump at the lower end of the trigram frequency distribution in the lower panel of **Figure 1**.

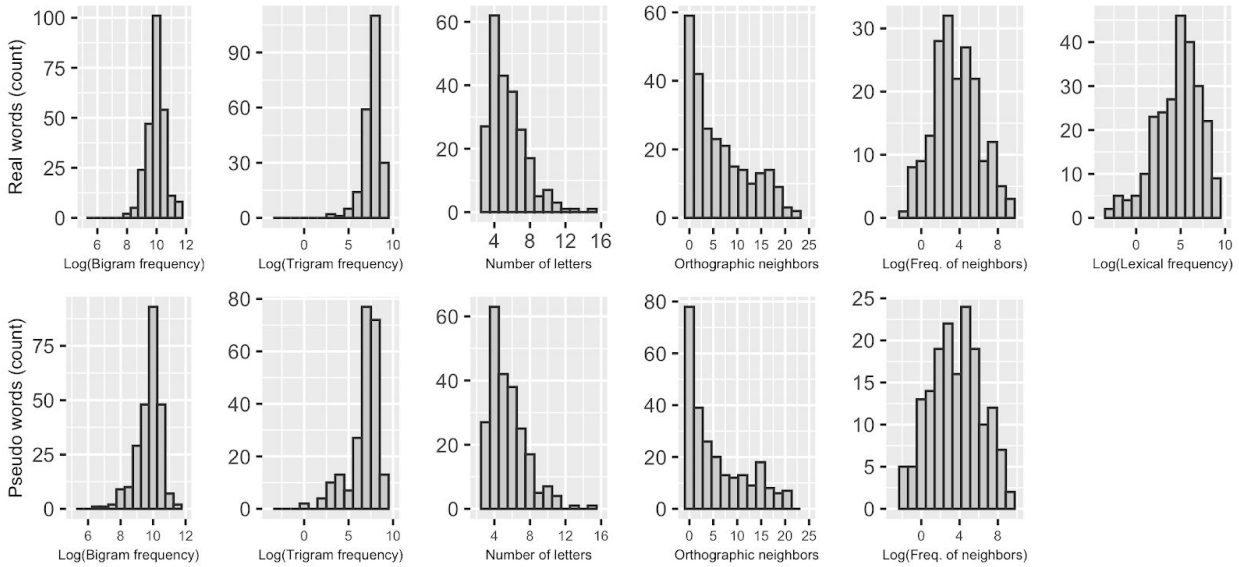

Figure 1. Lexical and orthographic properties of the stimuli. *(Top panel) Bigram frequency, trigram frequency, number of letters, number of orthographic neighbors, frequency of orthographic neighbors and lexical frequency for the 250 real words used in the lexical decision task (LDT). (Bottom panel) Bigram frequency, trigram frequency, number of letters, number of orthographic neighbors, frequency of orthographic neighbors for the 250 pseudo words used in the LDT.*

Participant recruitment and consent procedure

Participants were recruited from two research participant databases: (1) The Stanford University Reading & Dyslexia Research Program (http://dyslexia.stanford.edu) and (2) The University of Washington Reading & Dyslexia Research Program (http://ReadingAndDyslexia.com). Each database includes children and adults who have (1) enrolled and consented (and/or assented) to being part of a research participant pool, (2) filled out extensive questionnaires on demographics, education history, attitudes towards reading and history/diagnoses of learning disabilities, (3) been validated through a phone screening to ensure accuracy of basic demographic details entered in the database. Many of the children and adults have participated in one or more study at Stanford University or the University of Washington and, as part of these studies, have undergone an in-person assessment session including standardized measures of reading abilities (Woodcock-Johnson IV Tests of Achievement, Test of Word Reading Efficiency - 2), verbal abilities (Welschler Abbreviated Scales of Intelligence II (WASI-II), Vocabulary subtest) and general reasoning abilities (WASI-II, Matrix Reasoning

subtest). Scores on standardized tests were analyzed if the testing had occurred within 12 months of completing the LDT task (scores were adjusted based on the expected improvement since the time of testing).

Participants in the database were emailed a brief description of the study and a link to a digital consent form and the online LDT task. If participants provided consent, their LDT performance was linked to their data in the research database (questionnaires and standardized tests). Participants also had the option to complete the online experiment and remain anonymous. A total of 120 children and adults between the ages of 6.27 years and 29.15 (M = 11.61, SD = 3.67) participated in this study. These subjects spanned the full range of reading abilities as measured by the Woodcock Johnson Basic Reading Skills (WJ-BRS) standard scores (M = 97.7, SD = 18.1, min = 57, max= 142) and TOWRE Index standard scores (M = 92.4, SD = 18.9, min = 55, max= 134). Many subjects were recruited for studies on dyslexia and, based on the WJ-BRS, 25% of the subjects met a typical criterion for dyslexia (standard score more than 1 SD below the population mean). 37.5% of subjects were below the 1 SD cutoff in terms of TOWRE Index. We treat reading ability as a continuous measure throughout as opposed to imposing a threshold that defines a group of participants as individuals with dyslexia.

Complete demographic information for all the participants can be found at: https://github.com/yeatmanlab/ROAR-LDT/blob/master/data_allsubs/metadata_all.csv

Data analysis

All analysis code and data is publicly available at: https://github.com/yeatmanlab/ROAR-LDT, with a README file that documents how to reproduce each figure and statistic reported in the manuscript. Correct/incorrect responses and log transformed response times for each item were concatenated across subjects into a large table. Two subjects were identified as outliers and removed from further analysis based on the following procedure: First, median response times (RTs) were calculated for each participant. Then, participants were excluded if their median response time was more than 3 standard deviations below the sample mean. The two subjects who met this criterion performed the LDT at near chance accuracy level, suggesting that their extremely fast RTs were indicative of random guessing (lack of experiment compliance). These participants are displayed in the figures as open circles for reference. For the analysis of RT data, responses shorter than 0.2 seconds or longer than 5 seconds were removed. Then, quartiles and the interquartile range (IQR) of the RT distribution were calculated per participant. Responses that were longer than 3 times the IQR from the third quartile, or shorter than 3 times the IQR from the first quartile, were further excluded (see [47,48] for similar outlier exclusion procedures). These steps resulted in the exclusion of 4.6% of total trials. Analysis of RT data was conducted using the lme4 package in R [49].

STUDY 1: RESULTS

Lexical decision task performance and reading ability

To examine whether response times (RT) recorded in the online LDT replicated classic findings in the literature we fit linear mixed effects models to log transformed RT data for all correct trials. Each model included random intercepts for each word and subject. We examined five different models containing fixed effects of (1) lexical status (real word vs. pseudoword), (2) log transformed lexical frequency (real words only), (3) log transformed bigram frequency (for pseudowords only), (4) word length and (5) the interaction effect of word length and participant reading ability as measured with the WJ-Word-ID. As expected, we found: (1) RT was longer for pseudowords compared to real words (**β** = 0.11, SE = 0.00648, t = -17.245) (similar to [50]); (2) responses to real words were faster with increasing lexical frequency (**β** = -0.021, SE = 0.0013, t = -15.74) (similar to [50–52]); (3) responses to pseudowords were slower with increasing bigram frequency (**β** = 0.035, SE = 0.0057, t = 6.13) (similar to [50,53,54]); (4) responses to words and pseudowords were slower with increasing stimulus length (**β** = 0.032, SE = 0.0038, 3.3% increase per letter, t= 8.5) (similar to [50]); and (5) there was a significant length by reading ability interaction (**β** = -0.008, SE = 0.00195, t = 4.05) indicating that the effect of word length on reading speed diminishes as reading ability improves (similar to [43,50]).

Stimuli for the LDT were chosen to span a broad range of lexical and orthographic properties (**Figure 1**), with the aim of providing a sufficiently large difficulty range for detecting associations between accuracy on the LDT and reading ability measured with the WJ-Word-ID. Indeed, overall accuracy, calculated as the percent of correct responses on the 500 LDT trials, was highly correlated with WJ-Word-ID (**Figure 2 left panel**, r = 0.91, p < 0.00001, disattenuated r = 0.94). Accuracy for pseudowords (250 trials) had a slightly higher correlation with WJ-Word-ID (r = 0.86, p < 0.00001) than did accuracy for real words (r = 0.74, p < 0.00001), and this difference in correlation approached significance based on William's test (p= 0.08; **Figure 2 right panel**).

In contrast, median RT across correct responses in the LDT was not correlated with WJ-Word-ID (r = -0.06, p = 0.69). No correlations were found when median RT was calculated separately for words and pseudowords (r = -0.13, r = -0.08, respectively), even though RTs followed the expected patterns in other analyses (e.g., length effect, see preceding paragraph). However, when the RT analysis was limited to the subset of the subjects who performed above 70% correct (N=28 with WJ-Word-ID scores), a significant negative correlation was found between RT and WJ-Word-ID (r = -0.53, p < 0.005). The correlation between WJ-Word-ID and median RT for real words (r = -0.53, p < 0.005) was similar to the correlation with median RT for pseudowords (r = -0.52, p < 0.005), and this difference was not significant according to William's test (p = 0.79). Thus, response accuracy is a much stronger predictor of reading ability than RT (William's test t =-5.72, p < 0.00005), and RT is only informative in relatively high performing subjects.

The impact of participant characteristics on the LDT-WJ correlation was examined through mediation and moderation analyses. The relation between LDT accuracy and WJ-Word-ID raw scores was not moderated by WJ-Word-ID standard scores, age or number of

months since testing (t = 0.71, p= 0.48; t = 0.12, p= 0.91; t = 1.24, p = 0.22). These results were further confirmed by a mediation analysis (CI [-0.001, 0.00], p= 0.69; CI [-0.001, 0.00], p= 0.87; CI [-0.0002, 0.00], p = 0.49). Together, these analyses suggest that LDT accuracy is a good approximation of standard reading measures for children with reading profiles ranging from severely impaired to exceptional, and ages between 6 and 18 years. It also suggests that LDT accurately predicts reading ability measured within the past 12 months.

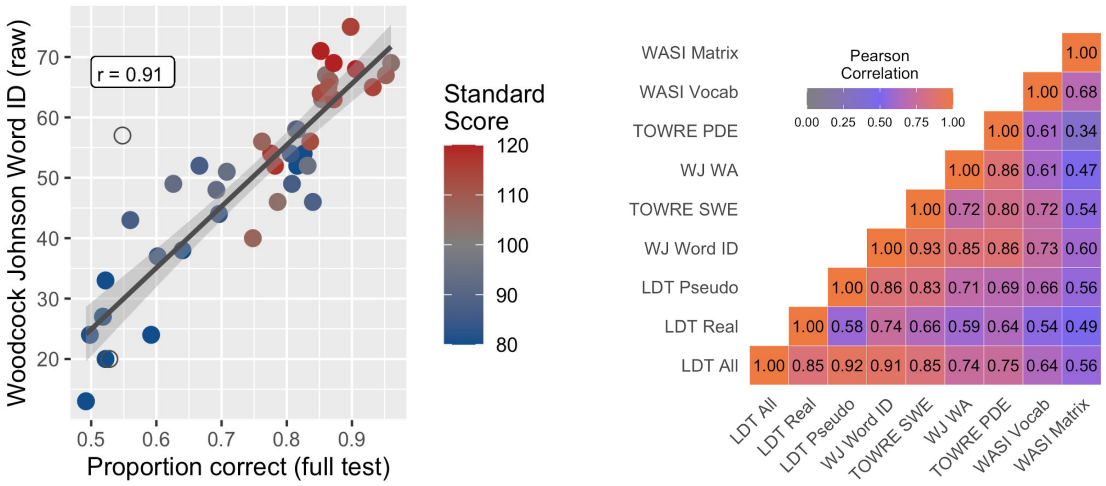

Figure 2. Accuracy on the lexical decision task predicts reading ability. *(Left panel) Scatter plot showing each participants' accuracy (proportion correct) on the 500 item LDT versus their raw score on the Woodcock-Johnson IV Letter Word Identification (WJ-Word-ID) test. Points are colored based on WJ-Word-ID standard scores, indicating that the relationship holds for struggling readers and those with dyslexia (dark blue), average readers (gray), above-average and exceptional readers (dark red). Open circles indicate outliers that were excluded from the analysis based on extremely fast response times on the LDT (see Methods). (Right panel) Correlation matrix between different measures of reading performance. The colormap represents the strength of correlation between each pair of variables. LDT All: overall accuracy on LDT; LDT Real: accuracy for real words; LDT Pseudo: accuracy for pseudowords. TOWRE SWE and PDE: Test of Word Reading Efficiency Sight Word Efficiency and Phonemic Decoding Efficiency; WJ WA: Woodcock Johnson Word Attack; WASI Vocab and Matrix: Wechsler Abbreviated Scales of Intelligence Vocabulary and Matrix Reasoning.*

Performance across different measures of reading ability is typically highly correlated. Indeed, different tests are frequently used interchangeably in different studies. For example, WJ-Word-ID scores and TOWRE-SWE scores are typically highly correlated (r=0.93 in our sample), even though the former is an untimed test while the latter is a timed test that combines real word reading accuracy and speed. In fact, it is common to use a threshold on either test to group subjects into dyslexic versus control groups in many research studies [21,55–57]. The average Pearson correlation between the timed (TOWRE), untimed (WJ), real word (Word-ID and SWE) and pseudoword (Word Attack and PDE) standardized reading measures was r = 0.83 (range: r = 0.72 to r = 0.93; **Figure 2 right panel**). Given the correlation between LDT accuracy and WJ-Word-ID, it is not surprising that LDT accuracy was also significantly correlated with all the standardized reading measures (**Figure 2 right panel**). The correlations between LDT accuracy

and Vocab (r = 0.64) and Matrix Reasoning (r = 0.56) were also nearly identical to the correlations between other reading measures and verbal abilities (r = 0.61 to 0.73) and reasoning abilities (r = 0.34 to 0.60, **Figure 2 right panel**). Thus, LDT response accuracy seems a suitable measure of reading ability as it behaves in a similar manner to the other established, standardized reading measures.

Item response theory analysis of responses to individual words and pseudowords

Next we turn to the challenge of constructing a LDT with suitable psychometric properties for applicability in a research setting. The word list used in Study 1 was constructed to broadly sample different types of words and pseudowords. In doing so, our goal was to use item response theory (IRT) to select the optimal subset of stimuli that measure reading ability reliably and efficiently across the full continuum of reading abilities. Our approach was to remove items that were: (1) not correlated with overall test performance and WJ-Word-ID performance, (2) not well fit by the Rasch model and (3) examine item disrcimination based on the two parameter logistic model (2PL). We then constructed an optimized set of three short stimulus lists, matched in terms of difficulty.

Step 1: Remove words that are not predictive of overall performance

As a first step, we computed the correlations between item responses, LDT accuracy calculated on the full 500 item test, and WJ-Word-ID scores. Performance on some stimuli, like the pseudoword "insows", was highly correlated with overall test performance ($r = 0.61$, $p < 0.00001$) and WJ-Word-ID score ($r = 0.60$, $p = 0.000035$). This means that a participant's response to that word was highly informative of their reading ability. Other stimuli, like the pseudoword "timelly" and the real words "napery" and "kind" provided no information about overall test performance, as indicated by correlations near (or below) zero. We removed 71 items that did not surpass a lenient threshold of r = 0.l. **Figure 3** shows words that were rejected (red) and retained (blue) based on this criterion.

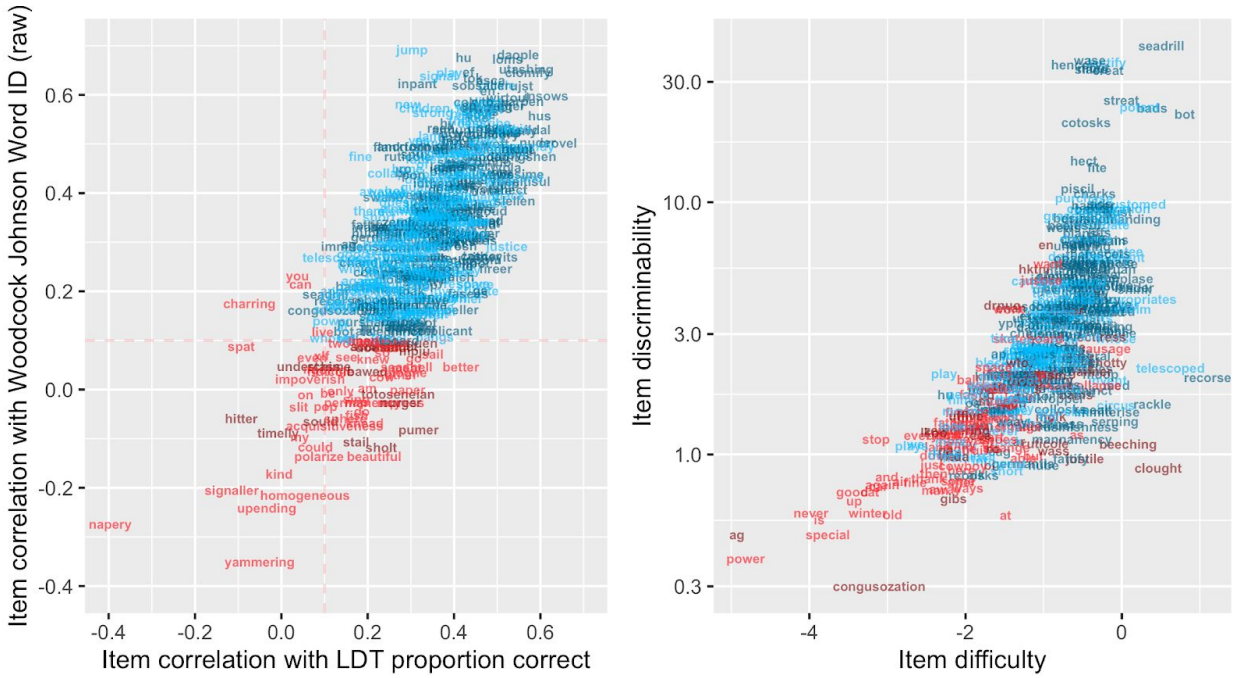

Figure 3. Item analysis. *(Left panel) For each stimulus, we plot the correlation between performance on this item and LDT accuracy (x-axis) as well as the correlation between performance on this item and WJ-Word-ID raw score (y-axis). Items that were not correlated (threshold r <= 0.10, Step 1) with overall test performance or WJ-Word-ID were removed (plotted in red). Items that passed the r>0.1 threshold are plotted in blue. (Right panel) Difficulty (x-axis) and discriminability (y-axis, based on the 2 parameter logistic model) for each item that survived the correlation threshold. Red items were those that were removed in step 2 or 3, based on poor fit statistics of the Rasch (Step 2) or 2 parameter logistic (Step 3) models. Pseudowords are depicted in darker colors, real words are depicted in lighter colors.*

Step 2: Remove words that are not well fit by the Rasch model

The Rasch model (1 parameter logistic with a guess rate fixed at 0.5) [58] was fit to the response data for the 429 items that remained after *Step 1* for all 118 subjects, using the MIRT package in R [59]. Items were removed if the infit or outfit statistics were outside the range of 0.6 to 1.4, a relatively lenient criterion intended to remove items with unpredictable responses [60]. The model was iteratively refitted until all items met this criterion. The process resulted in the removal of 129 items with response data that did not fit well with the predictions of the Rasch model (**Figure 3 right panel**, red items).

Step 3: Examining item discrimination with the two parameter logistic model

Next we fit the two parameter logistic (2PL) model (guess rate fixed at 0.5) to the response data, for the 300 items that were retained after *Step 2*. The 2PL model fits item response functions with a difficulty parameter (*b*) as well as a slope parameter (*a*); we sought to remove any remaining items that were inefficient at discriminating different levels of reading ability based on the slope of the item response function. The right panel of **Figure 3** shows the item difficulty

and discrimination parameters from the 2PL model. We removed 1 item with a low slope (a < 0.7). Of the remaining 299 items, 114 were real words and 185 were pseudowords.

To better understand the lexical and orthographic properties of the items that were selected for the final test, we compared the retained- versus rejected-items based on word length and log frequency of their orthographic neighbors (see Supplementary Figure 1). We found a significant effect for word length (Kruskall-Wallis H(3) = 26.6, $P < 0.0001$): Retained real words were significantly longer than rejected real words (p <0.0001) and retained pseudowords (p = 0.050). We expected that the log frequency of the orthographic neighbors would align with the word length effect reported above (i.e,. the final set of retained words would have less frequent orthographic neighbors compared to the rejected words). Our analysis confirmed this expectation (n = 359, Kruskall-Wallis H(3)=14.5, P=0.0023). Indeed, pairwise comparisons confirmed that retained real words had significantly less frequent orthographic neighbors compared to the rejected real words (p <0.0004). Therefore, real words, but not pseudowords, that are longer and have less frequent orthographic neighbors, are retained as a result of the IRT analysis, indicating their utility for measuring reading ability.

Step 4: Test optimization

Finally, we sought to create an efficient and reliable test that was composed of three short word-lists with the following properties: (a) quick to administer through the web-browser, (b) balanced in number of real words and pseudowords, (c) matched in terms of difficulty and (d) optimally informative across the range of reading abilities. Based on feedback from subjects in Study 1, we surmised that a list of 80 items (40 real words, 40 pseudowords) was a suitable length for maintaining focus. Thus, out of the 299 items that remained after Steps 1-3, we constructed three lists with items that were equated in difficulty based on b parameters (estimated with the Rasch model). We further ensured that items spanned the full difficulty range to maximize test information for the lowest and highest performing participants. Since only 114 real words were retained after the optimization procedure described above, lists were generated with 40 pseudowords and 38 real words. 2 new real words were then added to each list (described below in *Study 2*).

Figure 4 shows the test information function for each of the three lists. The composite of the three word lists (234 item; scores combined across lists) was extremely reliable with a reliability of 0.98. The composite LDT has a reliability comparable to those of the WJ and TOWRE composite indices: WJ Basic Reading Skills, average reliability = 0.95 [37], and TOWRE Total Word Reading Efficiency, average reliability = 0.96 [61]. Performance on the individual word lists was also highly reliable, with an ICC of 0.93 (**Figure 4 upper right**; about the same or higher than the individual TOWRE and WJ test forms: WJ Word Attack reliability = 0.90 and WJ Letter Word Identification had reliability = 0.94 [37]; TOWRE Sight Word Efficiency had an reliability = 0.91 and TOWRE Phonemic Decoding Efficiency reliability = 0.92 [61]. Performance on each individual word list was highly correlated with WJ-Word-ID scores (**Figure 4 bottom panel**; List A: r = 0.86; List B: r = 0.85; List C: r = 0.86) suggesting that administering a single 80-item lexical decision task (~2-3 minutes) would be a suitable measure of reading skill for some research purposes.

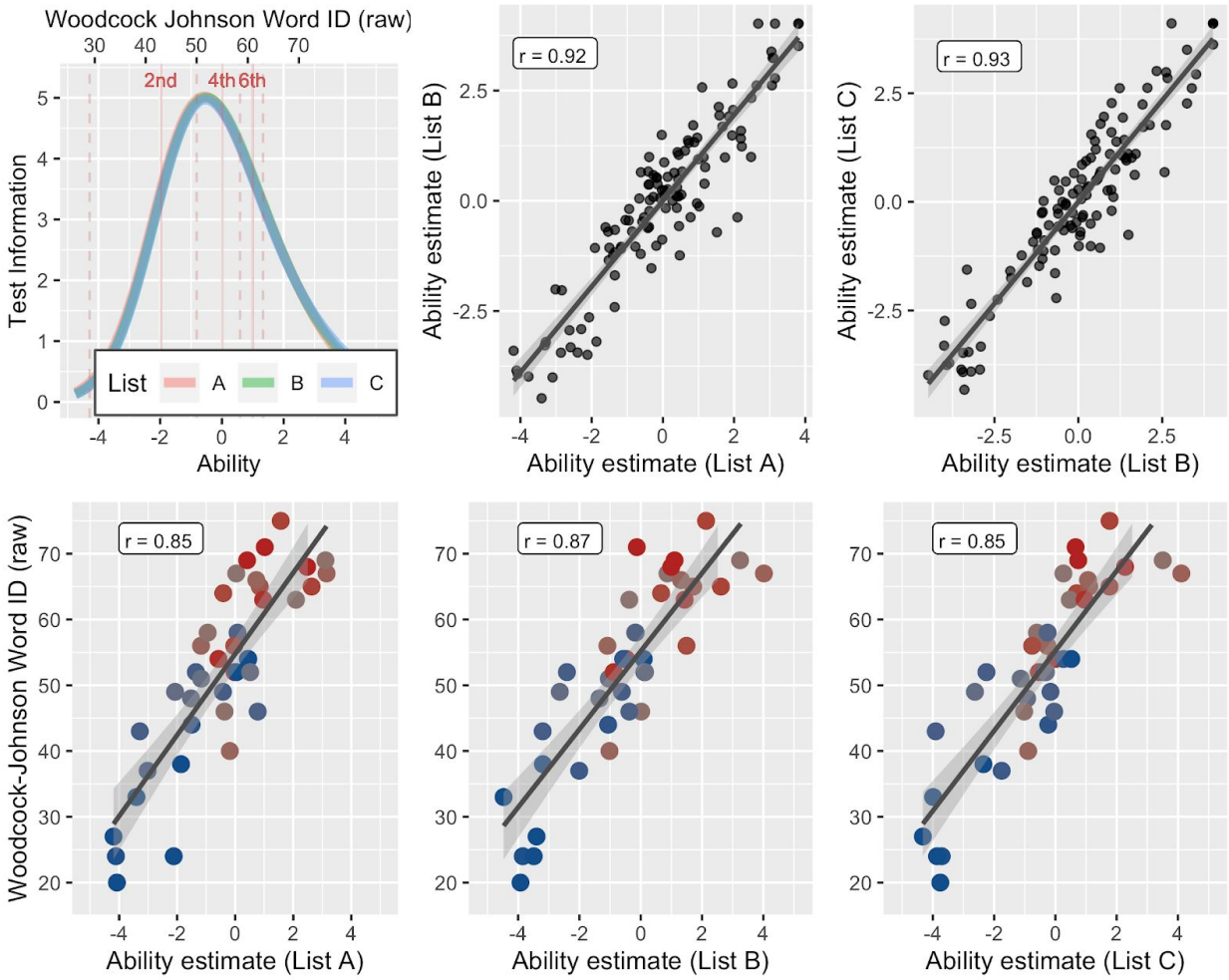

Figure 4. Optimized word lists. *(Top left) Test information functions for the three word lists. The curves are almost perfectly overlapping over the full range, predicting near identical performance on each list irrespective of individual ability. The x-axis shows ability estimates based on the Rasch model. The upper x-axis shows the estimated WJ-Word-ID raw score equivalent based on the linear relationship between ability estimates and WJ-Word-ID scores. The pink lines indicated grade level equivalents for WJ-Word-ID scores (based on the WJ manual). Test information is high for participants scoring above the 1st grade level on the WJ-Word-ID. (Top right) Scatter plots of ability estimates based on each list. Test-retest reliability for each individual list is greater than r=0.93. (Bottom panel) Correlations between ability estimates on each list and WJ-Word-ID scores. The blue-gray-red colormap indicates standard scores as in Figure 2. Even a single short list provides an accurate estimate of WJ-Word-ID ability.*

STUDY 1: DISCUSSION

The primary goal of Study 1 was to evaluate the suitability of a browser-based lexical decision task as a measure of reading ability. Lexical decision is commonly used to interrogate the mechanisms of word recognition, and previous studies have shown: (a) differences in task performance in dyslexia, (b) changes in task performance over development and (c)

relationships to various measures of reading ability. Thus, lexical decision was a strong candidate for an automated measure of reading ability. Results from Study 1 revealed that lexical-decision accuracy was a very accurate predictor of reading ability, measured with the WJ-Word-ID. Moreover, reliability analysis confirmed that the lexical decision task has excellent psychometric properties. Thus, by selecting the optimal list of words to span the continuum of reading skill, it might be possible to design a quick, valid and reliable measure of reading ability that could be administered to readers of all ages through the web-browser. We used IRT to select individual words with suitable measurement properties and created three short versions of the LDT that were matched in terms of difficulty. In the future, leveraging the IRT analysis to implement the LDT as a computer adaptive test would likely lead to even more efficient and accurate measures of reading ability. The goal of *Study 2* is to validate and norm the short-form LDT as a standardized measure of reading ability.

Through the process of optimizing the items to be included in the short-form test, we made four interesting observations that are worth noting (and all the data are publicly available for other researchers who may be interested in further item-level analyses). First, accuracy is a much better predictor of reading ability than RT (at least for children learning to read in English). Previous studies using RT have found statistically significant, but relatively weak relationships between RT and reading ability [44]. Indeed, this finding is replicated in our data (page 7). RT is likely a useful measure for participants at the high end of the reading continuum and, with a larger sample, might be useful for making more fine-grained discriminations of automaticity among skilled readers. Second, pseudowords are better at discriminating different levels of reading ability than are real words (as indicated by higher correlation with reading ability and more pseudowords being retained in the IRT analysis). It is well established that properties of the real words included in a LDT influence responses to pseudowords and vice versa [54,62,63]. Third, we investigated the orthographic properties of the items that were selected by the IRT analysis. Retained real words were longer and had lower frequency of orthographic neighbors than removed real words, an effect that was not observed for pseudowords (see Supplementary Figure 1). Finally, including items that span a wide range of difficulty (**Figure 3, 4**) is critical to a lexical-decision based measure of reading ability.

STUDY 2: METHODS

The following sections lay out our plan for a larger-scale validation and norming study. Data collection for *Study 2* is planned for September 2020.

Browser-based lexical decision task (version 2)

A shorter version of the browser-based LDT (240 trials in Study 2 vs. 500 trials in Study 1) was implemented based on the three, optimized 80-item lists that were designed in Study 1. Each list was presented in a block of trials and the order of the lists was randomized as were the order of the words within each list (but each word was not mixed across lists). Otherwise, all the details of the experiment were maintained from version 1. New words were selected based on grade

level text with specific considerations to choosing easy vocabulary words for 1st and 2nd grade students..

Participant recruitment

N>=600 children between 1st and 6th grade (n >= 100 per grade) will be recruited through partnerships with the local school districts. These data will be used to establish norms for the browser-based LDT. Out of the full sample, 150 subjects will be administered a battery of standardized tests over Zoom. This will allow us to equate our norms to the scale of other tests that are commonly used in research.

REFERENCES

1. Snowling, M. Dyslexia as a phonological deficit: Evidence and implications. *Child Adolesc. Ment. Health* (1998).

2. Wagner, R. K. & Torgesen, J. K. The Nature of Phonological Processing and Its Causal Role in the Acquisition of Reading Skills. *Psychol. Bull.* **101**, 192–212 (1987).

3. Vellutino, F. R., Fletcher, J. M., Snowling, M. J. & Scanlon, D. M. Specific reading disability (dyslexia): What have we learned in the past four decades? *J. Child Psychol. Psychiatry* **45**, 2–40 (2004).

4. Pennington, B. F. From single to multiple deficit models of developmental disorders. *Cognition* **101**, 385–413 (2006).

5. Pennington, B. F. *et al.* Individual Prediction of Dyslexia by Single vs. Multiple Deficit Models. *J. Abnorm. Psychol.* **121**, 212–224 (2012).

6. Vidyasagar, T. R. & Pammer, K. Dyslexia: a deficit in visuo-spatial attention, not in phonological processing. *Trends Cogn. Sci.* **14**, 57–63 (2010).

7. Goswami, U. Sensory theories of developmental dyslexia: three challenges for research. *Nat. Rev. Neurosci.* **16**, 43–54 (2015).

8. Ramus, F. Developmental dyslexia: Specific phonological deficit or general sensorimotor dysfunction? *Curr. Opin. Neurobiol.* **13**, 212–218 (2003).

9. Bosse, M. L., Tainturier, M. J. & Valdois, S. Developmental dyslexia: The visual attention span deficit hypothesis. *Cognition* **104**, 198–230 (2007).

10. Joo, S. J., Donnelly, P. M. & Yeatman, J. D. The causal relationship between dyslexia and motion perception reconsidered. *Sci. Rep.* **7**, (2017).

11. Joo, S. J., White, A. L., Strodtman, D. J. & Yeatman, J. D. Optimizing text for an individual's visual system: The contribution of visual crowding to reading difficulties. *Cortex* **103**, (2018).

12. Stein, J. F. The current status of the magnocellular theory of developmental dyslexia. *Neuropsychologia* **130**, 66–77 (2019).

13. Stein, J. F. & Walsh, V. To see but not to read; the magnocellular theory of dyslexia. *Trends Neurosci.* **20**, 147–152 (1997).

14. Lovegrove, W. J., Bowling, A., Badcock, D. & Blackwood, M. Specific reading disability: Differences in Contrast Sensitivity as a Function of Spatial Frequency. *Science* **210**, 439–440 (1980).

15. Hulme, C. The implausibility of low-level visual deficits as a cause of children's reading difficulties. *Cogn. Neuropsychol.* **5**, 369–374 (1988).

16. Flint, S. & Pammer, K. It is the egg, not the chicken; dorsal visual deficits present in

dyslexia are not present in illiterate adults. *Dyslexia* **0**, 1–15 (2018).

17. Skottun, B. C. The magnocellular deficit theory of dyslexia: The evidence from contrast sensitivity. *Vision Res.* **40**, 111–127 (2000).

18. Skottun, B. C. The need to differentiate the magnocellular system from the dorsal stream in connection with dyslexia. *Brain Cogn.* **95**, (2015).

19. Merzenich, M. M. *et al.* Temporal processing deficits of language-learning impaired children ameliorated by training. *Science* **271**, 77–81 (1996).

20. Tallal, P. Auditory temporal perception, phonics, and reading disabilities in children. *Brain Lang.* **9**, 182–198 (1980).

21. O'Brien, G. E., McCloy, D. R., Kubota, E. C. & Yeatman, J. D. Reading ability and phoneme categorization. *Sci. Rep.* (2018).

22. O'Brien, G. E., McCloy, D. R. & Yeatman, J. D. Categorical phoneme labeling in children with dyslexia does not depend on stimulus duration. *J. Acoust. Soc. Am.* **146**, 245–255 (2019).

23. Rosen, S. Auditory processing in dyslexia and specific language impairment: Is there a deficit? What is its nature? Does it explain anything? *J. Phon.* **31**, 509–527 (2003).

24. Lieder, I. *et al.* Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia. *Nat. Neurosci.* **22**, (2019).

25. Ahissar, M., Lubin, Y., Putter-Katz, H. & Banai, K. Dyslexia and the failure to form a perceptual anchor. *Nat. Neurosci.* **9**, 1558–1564 (2006).

26. Perrachione, T. K. *et al.* Dysfunction of Rapid Neural Adaptation in Dyslexia. *Neuron* **92**, 1383–1397 (2016).

27. Gabay, Y., Thiessen, E. D. & Holt, L. L. Impaired Statistical Learning in Developmental Dyslexia. *J. Speech Lang. Hear. Res.* **56**, 934–945 (2015).

28. Elleman, A. M., Steacy, L. M. & Compton, D. L. The Role of Statistical Learning in Word Reading and Spelling Development: More Questions than Answers. *Sci. Stud. Read.* **23**, 1–7 (2019).

29. Yarkoni, T. The generalizability crisis. *psyarxiv.com* (2020).

30. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).

31. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nature Human Behaviour* **1**, 1–9 (2017).

32. de Leeuw, J. R. jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behav. Res. Methods* **47**, 1–12 (2015).

33. Pinet, S. *et al.* Measuring sequences of keystrokes with jsPsych: Reliability of response times and interkeystroke intervals. *Behav. Res. Methods* **49**, 1163–1176 (2017).

34. Peirce, J. & MacAskill, M. *Building Experiments in PsychoPy*. (SAGE, 2018).

35. Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N. & Evershed, J. K. Gorilla in our midst: An online behavioral experiment builder. *Behav. Res. Methods* **52**, 388–407 (2020).

36. Bridges, D., Pitiot, A., MacAskill, M. R. & Peirce, J. The timing mega-study: comparing a range of experiment generators, both lab-based and online. (2020).

37. Schrank, F. A., McGrew, K. S., Mather, N., Wendling, B. J. & LaForte, E. M. Woodcock-Johnson IV Tests of Achievement. (2014).

38. Gershon, R. C. *et al.* NIH Toolbox for Assessment of Neurological and Behavioral Function. *Neurology* **80**, S2–S6 (2013).

39. Wilkinson, G. S. Wide range achievement test--revision 3. *Wilmington, DE: Jastak Association* **20**, (1993).

40. Wechsler, D. Wechsler individual achievement test--Second UK edition. *The Psychological Corporation* (2005).

41. Balota, D. A., Yap, M. J. & Cortese, M. J. *Visual Word Recognition: The Journey From Features to Meaning (A Travel Update)*. 285–375 (2006).

42. Seidenberg, M. S. & McClelland, J. L. A distributed, developmental model of word recognition and naming. *Psychol. Rev.* **96**, 523–568 (1989).

43. Martens, V. E. G. & de Jong, P. F. The effect of word length on lexical decision in dyslexic and normal reading children. *Brain Lang.* **98**, 140–149 (2006).

44. Katz, L. *et al.* What lexical decision and naming tell us about reading. *Read. Writ.* **25**, 1259–1282 (2012).

45. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203 (2019).

46. Keuleers, E. & Brysbaert, M. Wuggy: a multilingual pseudoword generator. *Behav. Res. Methods* **42**, 627–633 (2010).

47. Beyersmann, E. *et al.* Morpho-orthographic segmentation without semantics. *Psychon. Bull. Rev.* **23**, 533–539 (2016).

48. Ferrand, L. *et al.* Comparing word processing times in naming, lexical decision, and progressive demasking: evidence from chronolex. *Front. Psychol.* **2**, 306 (2011).

49. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models using lme4. (2014).

50. Di Filippo, G., de Luca, M., Judica, A., Spinelli, D. & Zoccolotti, P. Lexicality and stimulus length effects in Italian dyslexics: role of the overadditivity effect. *Child Neuropsychol.* **12**, 141–149 (2006).

51. Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H. & Yap, M. Visual word

recognition of single-syllable words. *J. Exp. Psychol. Gen.* **133**, 283–316 (2004).

52. Keuleers, E., Lacey, P., Rastle, K. & Brysbaert, M. The British Lexicon Project: lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behav. Res. Methods* **44**, 287–304 (2012).

53. Balota, D. A. Word frequncy, repetition, and lexicality effects in word recongition tasks: beyond measures of Central Tendency. *Journal of Experimental Psychology: General* vol. 128 32–55 (1999).

54. Ratcliff, R., McKoon, G. & Gomez, P. A Diffusion Model Account of the Lexical Decision Task. *Psychol. Rev.* **111**, 159–182 (2004).

55. Shaywitz, B. A. *et al.* Disruption of posterior brain systems for reading in children with developmental dyslexia. *Biol. Psychiatry* **52**, 101–110 (2002).

56. Kubota, E. C., Joo, S. J., Huber, E. & Yeatman, J. D. Word selectivity in high-level visual cortex and reading skill. *Dev. Cogn. Neurosci.* **36**, 100593 (2019).

57. Fletcher, J. M., Lyon, G. R., Fuchs, L. S. & Barnes, M. A. *Learning disabilities: From identification to intervention*. (Guilford PRess, 2006).

58. Rasch, G. *Probabilistic models for some intelligence and attainment tests*. (University of Chicago Press, 1980).

59. Chalmers, R. P. & Others. mirt: A multidimensional item response theory package for the R environment. *J. Stat. Softw.* **48**, 1–29 (2012).

60. Wu, M. & Adams, R. J. Properties of Rasch residual fit statistics. *J. Appl. Meas.* **14**, 339–355 (2013).

61. Torgesen, J. K., Wagner, R. & Rashotte, C. *TOWRE 2: Test of word reading efficiency*. (Pearson Clinical Assessment, 2011).

62. Rastle, K., Kinoshita, S., Lupker, S. J. & Coltheart, M. Cross-task strategic effects. *Mem. Cognit.* **31**, 867–876 (2003).

63. Kinoshita, S. & Mozer, M. C. How lexical decision is affected by recent experience: symmetric versus asymmetric frequency-blocking effects. *Mem. Cognit.* **34**, 726–742 (2006).

ACKNOWLEDGEMENTS

We would like to thank the Pavlovia and PsychoPy team for their support on the browser-based experiments. This work was funded by NIH NICHD R01HD09586101, research grants from Microsoft and Jacobs Foundation Research Fellowship to J.D.Y.

SUPPLEMENTARY MATERIAL

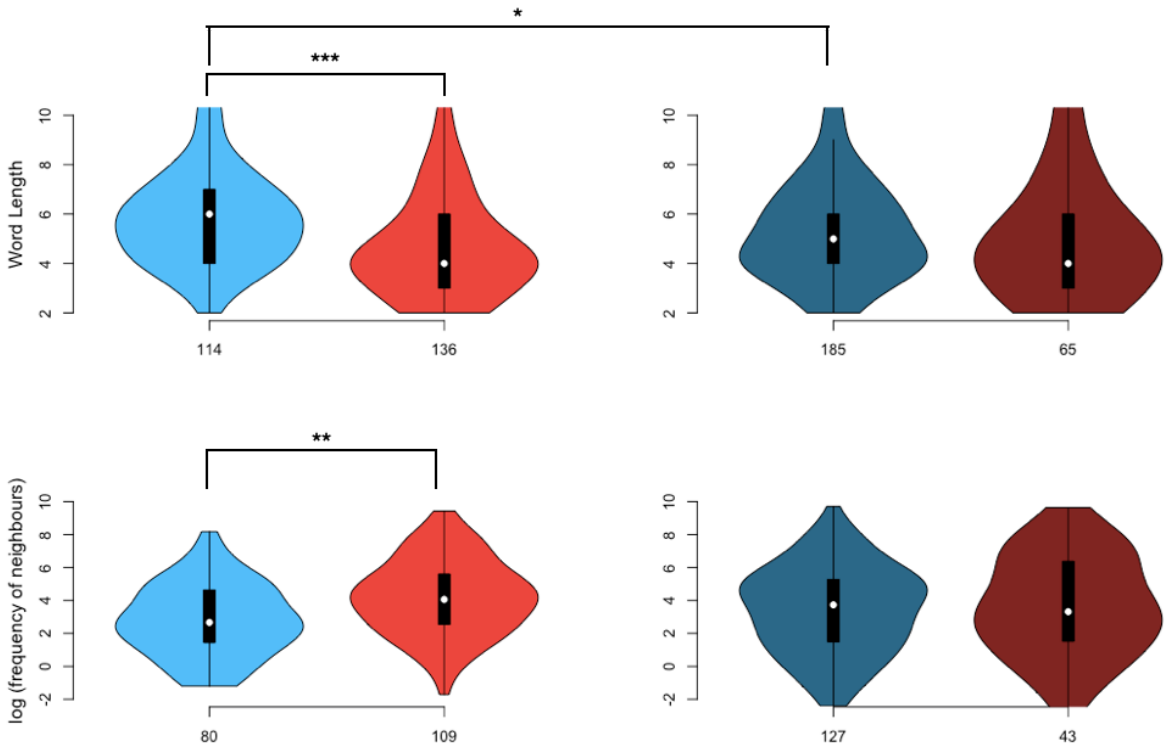

*Figure x: Top Panel (left: real words; right: pseudowords): shows the distribution of word length for the retained real words (BLUE) and rejected real words (RED) from step 3 of the IRT analysis. We found that the retained real words were longer than rejected real words (p <0.0001***) and retained pseudowords (p =0.050*). Bottom Panel (left: real words; right: pseudowords): Pairwise comparisons revealed that retained real words were significantly lower in log frequency of its orthographic neighbors compared to the rejected real words (p <0.0004**). These analyses show that the more unique a word is (longer and with low-frequency neighbors) the higher its discriminative power of reading performances. Darker colors are pseudowords and lighter colors are real words.*