Original Research

Hammill Institute on Disabilities

Assessment for Effective Intervention
1–11
© Hammill Institute on Disabilities 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/15345084211063533
aei.sagepub.com
SAGE

The Dyslexia Marker Test for Children: Development and Validation of a New Test

Trude Nergård-Nilssen, PhD[1] and Oddgeir Friborg, PhD[1]

Abstract

This article describes the development and psychometric properties of a new Dyslexia Marker Test for Children (Dysmate-C). The test was designed to identify Norwegian students who need special instructional attention. The computerized test includes measures of letter knowledge, phoneme awareness, rapid automatized naming, working memory, decoding, and spelling skills. Data were collected data from a sample of more than 1,100 students. Item response theory (IRT) was used for the psychometric evaluation, and principal component analysis for checking uni-dimensionality. IRT was further used to select and remove items, which significantly shortened the test battery without sacrificing reliability or discriminating ability. Cronbach's alphas ranged between .84 and .95. Validity was established by examining how well the Dysmate-C identified students already diagnosed with dyslexia. Logistic regression and receiver operating characteristic (ROC) curve analyses indicated good to excellent accuracy in separating children with dyslexia from typical children (area under curve [AUC] = .92). The Dysmate-C meets the standards for reliability and validity. The use of regression-based norms, voice-over instructions, easy scoring procedures, accurate timing, and automatic computation of scores, make the test a useful tool. It may be used in as part screening procedure, and as part of a diagnostic assessment. Limitations and practical implications are discussed.

Keywords

dyslexia, assessment, computerized, IRT

Dyslexia represents a difficulty in learning to decode print. Individuals with dyslexia often have problems because they do not master the critical factors underlying decoding and spelling, which, in turn, often hamper reading comprehension. To this date, many Norwegian children with dyslexia are missed due to the lack of norm-referenced assessment tools. A recent report reveals that 51% of the affected individuals are not identified or diagnosed until they enter high school (Solem, 2021). The purpose of this study was to develop a dyslexia marker test for Norwegian students and to examine its psychometric properties.

A considerable body of research shows that problems with orthography (Georgiou et al., 2021) and phonology (Snowling & Melby-Lervåg, 2016) are the major proximal causal risk markers for dyslexia. In alphabetic languages, learning to read starts with learning the mapping between phonemes (the smallest units of speech distinguishing one word from another) and graphemes (one letter, or a group of letters, that represent a speech sound, or phoneme). Difficulties with the ability to attend to, discriminate, and manipulate sounds in words are highly likely to lead to difficulties with mapping speech and print, or rather, mapping phonology and orthography. Family risk studies demonstrate that phonological problems are present long before formal reading instruction begins (Thompson et al., 2015).

It is widely reported however that phonological problems are neither necessary nor sufficient to account for dyslexia. Other known underlying risk factors include problems with learning letters (Torppa et al., 2016), rapid word retrieval (Parrila et al., 2020), and with working memory (Peng & Fuchs, 2014). The research literature shows that when they operate together with decoding and spelling problems, the risk markers accumulate toward a threshold for a diagnosis (Snowling et al., 2020). In a longitudinal study, Catts et al. (2017) found that children with a phonological awareness deficit in kindergarten were five times more likely to have dyslexia in second grade than children without such a deficit. This risk ratio substantially increased with the addition of deficits in both oral language and rapid naming. However, some of the children with heightened

[1]UiT The Arctic University of Norway, Tromso, Norway

Corresponding Author:
Trude Nergård-Nilssen, Department of Education, UiT the Arctic University of Norway, Universitetsvegen 39, Tromsø 9019, Norway.
Email: trude.nergard.nilssen@uit.no

risk were later found to be adequate readers, and Catts et al. (2017) argue for a multifactorial model of dyslexia which also includes protective factors that offset the impact of phonological and other cognitive-linguistic deficits. These ideas and similar findings are encapsulated in the multiple deficit model (MDM) originally proposed by Pennington (2006). The two fundamental tenets of the MDM is that multiple predictors contribute probabilistically to neurodevelopmental disorders (e.g., dyslexia) and that shared risk factors contribute to comorbidity. McGrath et al. (2020) highlight that the clinical and diagnostic implications of the MDM are that no single cognitive deficit can be used to rule in or out dyslexia at the individual level and that the dimensional and probabilistic nature of dyslexia (and other disorders) preclude clear mappings of cognitive profiles to the diagnosis. In line with these ideas, assessments should instead focus on the defining symptoms of dyslexia and should therefore include brief assessments of reading and their proximal skills (Snowling & Hulme, 2021).

The components underpinning reading performance and dyslexia appear universal. For example, Landerl et al. (2013) found that phoneme awareness and rapid automatized naming were strong concurrent predictors of developmental dyslexia across six European languages. A logistic regression analysis revealed however that more participants were classified correctly when the orthography was more complex. Similarly, Reis et al. (2020) report in their meta-analysis that orthographic transparency has a significant effect on the manifestation of dyslexia, with dyslexia symptoms being less marked and weaker in transparent compared to intermediate and opaque orthographies. Numerous studies furthermore show that in transparent languages, in which every grapheme roughly corresponds to one phoneme, reading accuracy hits the ceiling soon after formal reading instruction begins (Torppa et al., 2016). Nevertheless—although growth of reading skills is faster and follow a different trajectory in more regular orthographies than in English—phoneme awareness, letter-sound knowledge, and rapid automatized naming measured at the onset of literacy instruction are similarly important as predictors of variations in growth rate across languages (Caravolas et al., 2019). The Norwegian orthography, in which context this study took place, has consistent grapheme-phoneme correspondences (feed-forward consistency) but less consistent phoneme-grapheme correspondences (feedback consistency). Consequently, *spelling accuracy* is a bigger obstacle than reading accuracy to young readers and individuals with dyslexia, and similarly, *reading speed* appears to be a bigger obstacle than reading accuracy (Nergård-Nilssen & Hulme, 2014). The test reported here was designed to address these and other characteristic features of the Norwegian phonology and orthography by including a time-limited word decoding test and a spelling test that measures orthographic knowledge.

An increasing number of studies report high stability into adulthood and that weaknesses in phoneme awareness, rapid naming, and working memory are strong and residual correlates of dyslexia (Nergård-Nilssen & Hulme, 2014). At the same time, there is a growing body of literature affirming the value of providing early reading intervention to struggling readers. For example, Mathes et al. (2005) and Lovett et al. (2017) report that children who received intervention in first and second grade, made gains almost twice that of children receiving the same intervention in third grade and that the early intervention child continued to outperform the late intervention group. Miciak and Fletcher (2020) highlight that when risk for dyslexia is identified before Grade 3, the percentage of children who do not respond to explicit core and supplemental reading instruction are reduced to 2%–5%. It is thus critical to have valid tests for identifying this group of children available so that intervention can be provided to prevent or ameliorate reading disorders.

Assessment in all its forms—including screening, diagnostic testing, and monitoring—play a key role in any successful intervention. *Screening* can provide an indication of which children are "at risk" and would benefit from further support. A *diagnostic assessment*, on the contrary, can provide a clear indication of a child's strengths and weaknesses and specify which skills should be targeted within an intervention. It also gives a picture of the severity of the child's difficulties and to what extent support needs to be adapted. This study presents a norm-referenced test that is named The Dyslexia Marker Test for Children (acronym: Dysmate-C). The construction of the Dysmate-C was developed within the framework of the MDM. The defining markers are operationalized and construed as liabilities for dyslexia and include—in addition to decoding and spelling—letter knowledge, the ability to manipulate speech sounds (phoneme awareness), and the ability to name common symbols at speed (referred to as rapid automatized naming, or RAN). These markers are identified in cross-linguistic studies (Caravolas et al., 2019), in individual studies (Thompson et al., 2015), and in meta-analyses (Snowling & Melby-Lervåg, 2016). The Dysmate-C test was designed to identify children at risk for dyslexia and who thus need special instructional attention.

Ideally, the psychometric properties of any novel test are established by comparing how the results of the new test agree with the "true" outcome. In this study, some of our experimental instruments could not be validated against established measures of the same constructs. In the absence of a Norwegian "gold standard," we instead examined how well the Dysmate-C could identify children that were already diagnosed with dyslexia, and thus how well test outcomes would reflect our *a priori* expectations of poor performance in this group.

In summary, the main objective of this study was to develop a dyslexia marker test that can identify students who need special instructional attention. Another main objective was to establish its psychometric properties. A subordinate objective was to examine if gender, as a covariate, explains the ability (or trait score) on any of the tests. The research literature gives no reason to expect reading achievement or dyslexia to vary by gender. (Snowling & Hulme, 2021) point out that differences in the reported sex ratio between studies are likely related to measurement issues or to sampling bias. As the Dysmate-C is a new measurement, however, we wanted to rule out that it favors either gender.

Method

Sample

In Norway, children start school the calendar year they turn 6 years. Elementary school includes Year 1–7, and students transfer to secondary/high school the calendar year they turn 13. This study includes two samples: The first includes unscreened primary school children of which presence of reading disorders thus is unknown ($n > 1,000$). The second includes children that had been diagnosed with dyslexia prior to this study ($n = 50$). Boys and girls were roughly equally represented in both samples. Table S1 shows sample sizes reported for each subtest and grade level. All participants were recruited from 22 different schools across Norway. Students who were diagnosed with dyslexia were allocated to "the validation sample." Unfortunately, we do not have access to the diagnostic assessment that led to their diagnosis. As a rule, however, the school refers a student to the local educational-psychological service when dyslexia is suspected, that is, when a student shows signs of laborious reading of longer texts, or performs poorly on the national reading test in grade 5 or grade 8. The psychologist then typically carries out a standardized test that measures reading speed of connected text, reading comprehension, and listening comprehension, as well as reading-related skills. The diagnosis is established if certain criteria are fulfilled, for example, if orthographic word recognition is poor, or if listening comprehension is much better than reading comprehension.

The Dysmate-C Test

In this section, we will describe the original tests as they were administered to the children in the normative study. The Dysmate-C test battery is computerized and is presented on a laptop or tablet computer, with one available at each testing site. All instructions, practice items and corrective feedback are provided by voice-over. Feedback is provided in the practice sessions but not during the actual tests. Simple textual instructions and illustrations appear on the screen prior to each test to remind the child of the task requirements. The assessor is on hand to answer any additional questions during testing and to score oral responses. Responses are scored by the keyboard "A" (correct), "S" (wrong) or "D" (no response) keys, respectively, or alternatively by the left mouse button on the corresponding tab on the screen. The program automatically records scorings, and the timed tasks are regulated, and response time automatically recorded by a built-in timer. When assessor has entered the student's age and grade level, the test battery is composed accordingly. For example, the spelling test is not administered to Year-1 and -2 students, whereas the Letter Knowledge and Phoneme Isolation (PI) tests is considered too easy and thus not administered from Year 3 onwards. All except two tests (i.e., the letter knowledge and the spelling test) are either time-limited or speeded measures. For the youngest students, the session typically takes 10–15 minutes, whereas the session from Year 3 and above typically takes 20–30 minutes, including the spelling test.

The Letter Knowledge Test. There are 29 letters in the Norwegian alphabet. The child was asked to give the sounds and names of all letters. The letters were presented in random order to avoid use of rote learning from alphabet songs. If only the letter name was given, the child was prompted to provide the sound, and vice versa. Two points were awarded if the child produced a correct response for both the name and the sound, whereas one point was rewarded if only name or sound was produced correctly. This test was administered to students in grade 1 and 2, and to the validation sample.

Phoneme Isolation. To evaluate phoneme awareness, we asked the child to identify specific phonemes in words. In this test, the child was presented with sets of four illustrations and was asked to point at the object on the screen that either started or ended with a given sound (e.g., /s/). Four practice items introduced the test to familiarize the child to point at an object that "begins with" or "ends with" a sound. If no response was given within 10 seconds, the child was automatically presented with the next set and target sound. The score here was the number of correctly identified target words, with a maximum score of 16.

Phoneme Deletion. To measure phoneme awareness in students in Grade 3–7 further, we developed a phoneme deletion test. Here, the child was asked to produce the word that remained when a particular sound was omitted (e.g., /b/ in brød [bread], where rød [red] is the correct answer). Practice items introduced the test to familiarize the child with omitting the first, the last sound, or the middle sound, respectively. If no response was given within 10 s, the child was presented with the next word and target sound. The score here was the number of correctly produced words, with a maximum of 16.

Rapid Automatized Naming. To evaluate word retrieval speed and the ability to process information rapidly and automatically without effort, we developed a rapid naming task. Here, the child was asked to name different combinations of five colors and five objects as quickly and accurately as possible. All objects are highly familiar 2-syllable Norwegian words (nouns) that children know well before school entry. The objects were randomly displayed in one of the five colors, with each object and each color being presented 10 times. Three practice items introduced the test—the first was to name the colors, the second to name the objects, whereas the third was to name color-object combinations as quickly as possible. The score here was the time it took (in seconds) to name all 50 color-object items and for the number of errors made. If a child made a self-correction, the corrected response canceled out the incorrect response. The raw score here is the number of correctly named objects divided by the total naming time in seconds. This test was administered to all students.

Working memory. To tap working memory, we developed two sets of items—one consisting of words and one consisting of digits. In both tasks, the child listened to a sequence of items and was instructed to repeat them in reverse order. The backward digit sequence was completed prior to the backward word sequence. The words used in the present test were highly familiar Norwegian nouns. In both parts, the length of each sequence increased as the child responded correctly. Testing was terminated when the child failed two consecutive trials. Each correct response (i.e., a series of words or digits) was awarded one point, with a maximum of 12 for each subseries and 24 for the total score. This test was administered to all students.

Decoding. To evaluate decoding skills, we developed "The One-Minute Word Decoding Test" (hereafter named "the 1-min test" interchangeably). The test consisted of a list of 160 high-frequency words with increasing length and complexity. The child was asked to read aloud as many words, as accurately as possible, within a time limit of 60 s. The score here was the number of words read correctly in 1 min. This test was administered to all students.

Spelling. To evaluate orthographic skills, we developed "The Spelling Test" which consisted of 40 common words, with increasing length and complexity. To address orthographic knowledge, the test included words with silent letters and homophone words. All words were framed within a sentence to ensure the correct meaning, and then the target word was repeated. The child was asked to write the target word on the computer. Spelling and grammar checkers were disabled by the program. There were no time limits, and if necessary, the sentence and target word could be repeated once, (i.e., students were instructed to press the tab next to the blanks if repetition by voice-over was needed). Students used headphones on this test, and the test was thus administered to groups, instead of individually, which saved time. To avoid fatigue, the spelling test was automatically discontinued after five consecutive spelling errors. The score was the number of correctly spelled items, with a maximum score of 40. This test was administered only to students from grade 3–7.

Procedures

After approval by the principal, all children in that school were invited to the study. All parents received a letter containing relevant information concerning the test, a short questionnaire, and a self-addressed envelope. Parents of more than 1,100 students across grade 1–7 gave written consent to participate in the study.

All children were tested individually either in a quiet room at school, in a laboratory at the university, or at the clinic by research assistants who had received extensive training for the scoring procedures. All tests were administered in a single session, but research assistants were instructed to take breaks if a student showed signs of fatigue.

Data Analysis

The test construction phase generated a surplus of items, and consequently, a substantial portion of items might either be redundant (overlapping with other items), correlate unsatisfactorily with the dyslexia trait score (low discrimination), or be limited by ceiling or floor effects. For the present purpose, we used item response theory (IRT) as the basic analytical strategy to shorten the tests by removing redundant items. IRT has several advantages beyond Classical Test Theory (CTT) approaches such as factor analysis which estimates the simple correlation between an item and the latent factor. IRT instead estimates the probability that a person will make a correct response given their latent trait or ability level. It provides test item information that are less sample dependent, thus tolerating use of less representative samples, if necessary. To achieve sample invariant IRT parameters; however, a large and heterogeneous sample that is representative of as much of the population diversity in question—that is, the latent trait—is required (Cappelleri et al., 2014; Hambleton & Jones, 1993). We therefore conducted all IRT analyses on each Dysmate-C subtest using the entire sample. From the IRT parameters, item characteristic curves (ICCs) can be drawn that are highly valuable for the process of selecting the most informative items that discriminate between children across the entire latent trait continuum rather than just around the middle point (Hambleton & Jones, 1993). It can also quantify to what extent guessing underlies a correct response. These parameters are estimated independently from the ability estimation of the examinee, which benefits generalizability and interchangeability (Baker, 2001).

We also examined if differential item functioning (DIF) was a problem, which is present if different subsamples with the same estimated ability level have a different probability of responding correctly. Using gender as an example, this would occur if gender significantly predicted a different probability of a correct response for girls and boys having the same latent ability. As DIF items complicate the interpretation of examinees' true ability scores, such items are generally discarded. Finally, we examined the construct validity of the Dysmate-C by examining if the Dysmate-C trait scores correlated positively with age, and most importantly, with a variable defining a prior dyslexia diagnosis.

Data recording was conducted digitally, and responses were transferred to a secured database that was later copied without conversion or loss to Mplus. The Mplus software (version 8.4, Muthén & Muthén, 2017) was used for all psychometric analyses. We conducted an item factor analysis using the maximum likelihood estimator and a logit link function, which Mplus automatically converts to comparable IRT *b*- and *a*-parameters (for a 2PL model). The *b* parameter indicates at which ability or trait level (i.e., *theta*) approximately 50% of the examinees answer correctly. As the theta parameter has a mean of 0 and a standard deviation of 1, increasingly negative and positive *b*-values represent increasingly easy and harder items, respectively. The discrimination parameter (*a*) mimics the well-known factor loading in classical factor analyses through the formula $\lambda = a / \sqrt{1+a^2}$. Hence, an *a*-value of 1.0 corresponds to a factor loading of .71, which is generally considered good. The all-important uni-dimensionality assumption of IRT was tested using principal component analysis (PCA).

IRT may also estimate the degree of guessing taking place, which is a disadvantage for items using multiple-choice response options. For this purpose, we fitted a 3PL model that add a threshold (or *c*) parameter estimating the lower asymptote (or, guessing) at which even the least able child answers correctly. The *c* parameter may be estimated equal or free for all items, depending on which choice improves model fit. The model priors for the guessing parameters were set to *N*(1.386,1), as suggested by Muthén and Muthén (2017) for four-choice multiple response options. As the increasing parameterized IRT models (1PL, 2PL, and 3PL) represent nested models (e.g., 1PL being nested within 2PL, and 2PL within 3PL), we used the chi-square difference test to compare these models (Nguyen et al., 2014). We also examined if the Bayesian Information Criteria (BIC) index decreased by specifying a more complex model.

MIMIC modeling was conducted to examine if DIF was present for gender and age. Uniform DIF was present if the difficulty parameter (the *b*) for a particular item was significantly different depending on gender and/or age after accounting for the latent trait variable. Nonuniform DIF was present if the *b* parameter in addition varied as a function of the latent trait score. Uniform DIF was examined item by item by regressing both gender and age on the trait score in addition to the item in question. Nonuniform DIF was tested by adding the gender/age by latent trait interaction using XWITH in Mplus. Given the large number of item tests, an alpha level of at least <.01 was required for further consideration.

Selection of items. As the Dysmate-C was operationalized to detect liability for dyslexia, the most discriminatory items in the lower latent trait score area were deliberately retained. The procedure for selecting the best items was based on the item parameters, which was conducted separately for each test the Dysmate-C battery. First, we sorted items ascendingly in terms of their difficulty (*b*) parameters, which quickly identified clusters of items with equal difficulty. Within these clusters, items with the lowest discrimination (*a*) parameter were discarded, thus retaining the most discriminative items across the entire range of the difficulty area. In addition, items with high guessing values were discarded if a 3PL model was used. Finally, we plotted the test information curve (TIC) which represents the aggregate of all the item information curve plots. The TIC displays at which theta-score range examinees are best discriminated, and ideally the form of TIC should be as flat and uniform as possible, but in our case, it should yield more information at lower and middle (but not higher) ability levels to differentiate individuals with dyslexia from typical readers.

Reliability. Thissen (2000) suggests using the IRT test information index as an indication of composite reliability, which may be calculated from the standard error of the estimated theta-scores ($SE_{theta} = 1 / sqrt\left(information\right)$). Since these *SE* values are expressed on a standardized scale ($M = 0$, $SD = 1$), the reliability becomes $rel = 1 - SE_{theta}{}^{2}$. Thus, information values of 4, 5, and 10 mimics conventional reliability coefficients of .75, .80, and .90. We also provide classical lower-bound reliability estimates in terms of Cronbach's alpha.

Validity. The validity of the subscale test scores were examined in a series of logistic regression analyses with diagnostic status as a binary outcome and the subscale test scores as predictors, which were entered as standardized residual *Z*-scores (Z_{resid}). The use of regression-based norms has become standard practice within cognitive, neurocognitive, and the psychometric test psychology literature due to its advantages over the classical method of creating subgroups based on norm-relevant covariates (Magnusdottir et al., 2019). Accordingly, Z_{resid} scores were created using linear regression with the theta score as the outcome, and Year and Year-squared as predictors, to produce norm-based test scores (nonsignificant predictors were removed). This equation thus produced regression-based predicted theta scores that were subtracted from each child's observed theta score. By dividing this difference with the standard deviation of all residual scores, we obtained Z_{resid} scores that were used as predictors of diagnostic status. A negative or positive Z_{resid} score quantifies how much better or more poorly the child performs from what is expected given the child's age/year, and/or gender. The logistic regression analysis estimates group membership probability and provides an estimate of diagnostic accuracy.

To further evaluate the Dysmate-C's diagnostic accuracy, we conducted an receiver operating characteristic (ROC) curve analysis. This analysis uses a gold standard variable that truly defines the diagnostic status of the child (0 = "typical," 1 = "dyslexia"), and plots the true positive to the false-positive classification rate for all possible threshold values for the new test (the Dysmate-C). The overall accuracy is indexed by the area under curve (AUC) index (where AUC = 1.00 is perfect, whereas an AUC = .50 represents no discrimination). We added the Youden's index to help identify cut-off values that maximize the sensitivity and specificity of the new test.

Results

As part of the test construction phase, many items were generated as indicators of the six domains described above. The validation sample was included in the IRT analyses to estimate the IRT parameters based on the entire latent trait continuum. Please recall that three of the scales, the 1-Minute Word Decoding Test, the Rapid Automatized Naming test, and the Working Memory test, did not undergo IRT analyses as these are timed and recorded as a simple count of correct scores.

IRT and PCA

The Letter Knowledge Test. A PCA based on all 29 Letter Knowledge Test (LKN) items extracted a first eigenvalue of 25.22 (R^2 = 87%), whereas the second was 0.76 (additional R^2 = 2.6%), which supported uni-dimensionality clearly. The chi-square difference test was significant ($\chi^2_{\text{df=28}} = 379.71$, $p < .001$) and the BIC was substantially lower (ΔBIC = −188.42) in favor of the 2PL model. The item discrimination and difficulty parameters are presented in Table S2.

Items with low discrimination values within clusters of items with overlapping difficulty values were discarded, which amounted to 17 items. Table S2 provides the IRT parameters for the final 13-item scale. The ICCs for the final LKN scale indicated a strong discrimination ability, as well as adequate spread in the lower latent trait score area. As Figure S2 shows, the total information curve (TIC) based on all 29 and the final 13 items shows that the discrimination in the lower latent trait continuum area was well maintained by the reduced item set. In the item selection process, item *U* had high loading but significant uniform and non-uniform DIF with age and was replaced by items in proximity without DIF. Neither of the final selected 13 items exhibited significant DIF regarding gender nor age. The internal consistency (Cronbach's alpha) was .95. The TIC curve indicated high reliability (information >10; reliability >.90) in the theta-score area between −1.85 and −0.20, and acceptable reliability (>.75) between −2.17 and 0.09.

Phoneme Isolation. A PCA of the 16 PI items extracted a first eigenvalue of 9.68 (R^2 = 60.5%), whereas the second was 1.19 (additional R^2 = 7.4%). The correlation between the first and the second component was very high (r = .80), which together, strongly indicates uni-dimensionality. As guessing was possible due to the multiple-choice four-categorical response option, a 3PL IRT model was also tested. First, the 2PL model was preferred above the 1PL model due to a consistent reduction in the log-likelihood and the BIC (difference $\chi^2_{\text{df=15}} = 118.98$, $p < .001$; ΔBIC = −12.88).

The 3PL model with individual guessing parameters had worse model fit than the 2PL model (higher −2LL = 47.48 and higher BIC = 106.65), which also was the case for a 3PL model with the guessing parameter equal for all items (higher −2LL = 10.65 and higher BIC = 17.73). We thus discarded guessing in the further evaluation of these items.

The item selection process was similar as for the LKN by discarding lower discriminatory items among items with comparable difficulty parameters. The easiest item ("is" [ice]) was retained despite significant uniform DIF with age (centered at age of 8) as it was the most important item for discrimination in the lowest latent trait score continuum. Among the two hardest items ("kopp" [mug] and "fot" [foot]), which both had significant uniform DIF with age, "kopp" was retained to keep a single item providing discrimination in the highest trait score continuum. Adjustment for DIF reduced the discriminatory power of "kopp," but most importantly, increased the discrimination by "is."

Table S3 provides the IRT parameters for the final 10-item PI scale. Overall, the items indicate strong discrimination and particularly good spread in the lower latent trait score area. An EFA on these 10 items revealed a first and second eigenvalue of 6.75 and 0.73, and hence clearly supporting uni-dimensionality. As can be seen from Figure S3, the TIC based on all 16 and the final 10 items shows that the discrimination in the lower latent trait continuum area was adequately preserved compared to the full item version. The Cronbach's alpha was .84. The TIC curve indicated high reliability (>.90) in the theta-area between −2.05 and −0.83, and acceptable reliability (>.75) between −2.63 and −0.18.

Phoneme Deletion. A PCA of the 16 PD items extracted a single eigenvalue above 1 ($\lambda = 15.20$, $R^2 = 95.0\%$). The IRT 2PL model was better than the 1PL model in terms of a significant reduction in log-likelihood (difference $\chi^2_{df=15} = 215.37$, $p < .001$), as well as in ΔBIC = −109.35, and the 3PL model was clearly worse than the 2PL in terms of worse log-likelihood as well as BIC. Using the same item selection procedure, we retained 10 PD items. The item discrimination and location parameters are presented in Table S4, whereas the TIC for the full PD and the Final 11-Item Scale are displayed in Figure S4.

The item and TICs based on all 16 and the final 10 items shows that the discrimination in the lower latent trait continuum area was more restricted in the PD compared to the PI test. The TIC curve indicates high reliability (>.90) in a narrower theta-area as compared to PI, that is, between −1.16 and 0.43. If accepting lower reliability (>.75), the discrimination area widened to between −1.48 and 0.77. The lower-bound Cronbach's alpha was .93. Neither of the items showed significant DIF regarding gender or age.

Word spelling. A PCA of the 40 items extracted six eigenvalues above 1, with a clear deceleration at the fourth. As the first and second eigenvalues ($\lambda = 25.06$ and 3.08) accounted for 62.65% and 7.7% of the total item variance, and all geomin rotated loadings for the third component (if extracted) were below <.40, the three-component solution was discarded. Items 1–6, 8, and 9 correlated more strongly on a separate component according to the PCA. Specifying a multidimensional two-factor IRT model based on the PCA solution indicated a correlation of 0.93 between the two latent trait scores. As the standardized scaling differences between the two theta scores were of negligible magnitude ($SD = .16$), the overlap in trait scores was substantial, and all items were treated as uni-dimensional in the subsequent analyses.

The 2PL model was clearly better than the 1PL model according to the chi-square difference and BIC tests ($\chi^2_{df=39} = 523.27$, $p < .001$; ΔBIC = −282.90). Items with poorest discrimination properties among items with overlapping difficulty parameters were removed. As discrimination in the lower latent trait continuum was preferred, more of these items were retained. Table S5 shows that some of the 40 items showed significant DIF, but none of the selected 11 final items showed DIF regarding gender or age. The ICC curves show good spread of items across a wide trait area, even in the higher area which may be useful to identify good spellers. As Figure S5 shows, the TIC curve shows that the information value (construct variance) is quite well accounted for in the most important trait score area (between −2.5 and −1) despite the huge reduction in items. A PCA on these 11 items revealed a first and second eigenvalue of 8.69 and 0.86 and hence clearly supporting uni-dimensionality. The Cronbach's alpha was .89. The TIC curve indicated high reliability (>.90) in the theta-area between −2.14 and −0.35, and acceptable reliability (>.75) between −2.44 and 0.68.

Estimation of Regression-Based Norm Scores for Each Test Domain

Table 1 shows regression-based norm scores for the different Dysmate-C subscales (see method section describing the procedure). Beta coefficients showed that gender significantly predicted performance on the word spelling test only (with girls doing better than boys) and that increasing age predicted better performance across all tests, as expected. The age effect was in addition nonlinear as the performance gain decelerated around Year 5 or 6. As Table 1 shows, the Z_{resid} scores were strikingly lower for the children with dyslexia compared to the unaffected group on all subscales, except the LKN, PI, and working memory.

Analysis of Construct Validation: Logistic Regression and ROC Curve Analysis

Next, we conducted logistic regression analyses with dyslexia diagnosis as dependent variable and with Z_{resid} test scores as single (crude) and adjusted predictors (controlling for year and gender), respectively. Results showed that the probability of receiving a dyslexia diagnosis increased significantly with age, whereas gender was unimportant in this regard. As Table 2 shows, poorer test scores on the LKN, phoneme deletion, word spelling, rapid automatized naming, working memory, and on the 1-Minute Word Decoding Test significantly increased the odds of a diagnosis. The ROC analyses showed that AUC values were higher for four subscales: the 1-Minute Test, word spelling, rapid automatized naming and phoneme deletion, in falling order. Finally, we combined subscales with the highest AUC values to examine whether any combinations would improve the AUC. That is, we combined word spelling and rapid naming, and the 1-Minute Word Decoding Test and word spelling test. Although the precision improved slightly, combinations did not significantly improve the AUC estimate.

Pearson's correlation coefficients between the Dysmate-C subtests were all positive, ranging between .10 and 52 (average $r = .24$). Subjecting the Z_{resid} scores to a PCA extracted two components with an eigenvalue >1, revealing that the LKN and PI formed a separate cluster apart from the other tests. These two test scores also failed

Table 1. The Dysmate-C's Ability to Discriminate Between Typical Children and Children Diagnosed With Dyslexia.

Statistical values/ parameters	LKN	PI	PD	WS	RAN	WM	1-Min
N (typical/dyslexia)	748/25	737/51	991/26	449/50	427/50	680/51	667/51
Regression-based norms							
Intercept	.536***	.403***	.463	−.248***	.494***	7.840***	59.452***
Beta $_{\text{year}}$	.134***	.094***	.217***	.361***	.038***	.444***	8.584***
Beta $_{\text{year squared}}$	−.041***	−.043***	−.047***	−.055*	−.006***	Ns	−1.430***
Beta $_{\text{gender}}$	ns	ns	Ns	−.155*	ns	Ns	ns
Z_{resid} scores (*M*, *SD*)							
Normal	.01 (.98)	.01 (1.01)	.02 (1.00)	.14 (.92)	.14 (.94)	.06 (1.01)	.10 (.92)
Dyslexia diagnosis	−.27 (.82)	−.16 (.78)	−.82 (1.00)	−1.14 (.89)	−.97 (.87)	−.32 (1.01)	−1.46 (.67)

Note. LKN = letter knowledge; PI = phoneme isolation; PD = phoneme isolation; WS = word spelling; RAN = rapid automatized naming; WM = working memory; 1-Min = 1-Minute Word Decoding; Z_{resid} = Standardized deviation from expected norm-referenced value.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 2. Logistic Regression and ROC Curve Analysis.

Statistical parameters	Letter knowledge	Phoneme isolation	Phoneme deletion	Word spelling	Rapid naming	Working memory	The 1-min decoding test
Logistic regression							
Crude OR$_{\text{theta score}}$	$.77_{.52\text{–}1.13}$	$.86_{.66\text{–}1.11}$	$.44_{.29\text{–}.65}$	$.26_{.18\text{–}.38}$	$.20_{.12\text{–}.32}$	$.69_{.51\text{–}.94}$	$.11_{.06\text{–}.18}$
Adj OR$_{\text{theta}}$	$.46_{.25\text{–}0.86}$	$.78_{.57\text{–}1.09}$	$.41_{.27\text{–}.62}$	$.23_{.16\text{–}.35}$	$.22_{.14\text{–}.37}$	$.71_{.52\text{–}.96}$	$.09_{.05\text{–}.17}$
OR$_{\text{year}}$	$11.96_{1.58\text{–}90.25}$	$3.91_{1.83\text{–}8.33}$	$3.43_{1.35\text{–}8.73}$	$8.86_{2.42\text{–}32.45}$	$7.30_{1.86\text{–}28.62}$	$4.25_{1.72\text{–}10.51}$	$6.20_{2.07\text{–}18.60}$
OR$_{\text{year squared}}$	$.61_{.36\text{–}1.03}$	$.79_{.63\text{–}.99}$	$.77_{.57\text{–}1.04}$	$.59_{.40\text{–}.85}$	$.66_{.45\text{–}.96}$	$.72_{.56\text{–}.94}$	$.71_{.51\text{–}.98}$
OR$_{\text{gender}}$	$1.08_{.46\text{–}2.52}$	$1.12_{.61\text{–}2.04}$	$.95_{.42\text{–}2.18}$	$1.13_{.56\text{–}2.27}$	$1.32_{.62\text{–}2.80}$	$1.01_{.55\text{–}1.83}$	$1.33_{.56\text{–}2.71}$
ROC curve							
AUC theta$_{95\%\ \text{CI}}$	$.577_{.492\text{–}.663}$	$.588_{.523\text{–}.654}$	$.724_{.627\text{–}.820}$	$.850_{.805\text{–}.896}$	$.812_{.750\text{–}.873}$	$.601_{.516\text{–}.686}$	$.922_{.891\text{–}.952}$
Optimal Z_{resid} cut-off	−1.547	−.559	$-.951^{\text{sens}+}$ $-.168^{\text{spec}+}$	$-.686^{\text{sens}+}$ $-.472^{\text{spec}+}$	$-.700^{\text{sens}+}$ $-.533^{\text{spec}+}$	$-.984^{\text{sens}+}$ $-.295^{\text{spec}+}$	−.729
Sensitivity$_{95\%\ \text{CI}}$	.929	.725	$.839^{\text{sens}+}$ $.581^{\text{spec}+}$	$.820^{\text{sens}+}$ $.758^{\text{spec}+}$	$.820^{\text{sens}+}$ $.761^{\text{spec}+}$	$.856^{\text{sens}+}$ $.635^{\text{spec}+}$	.816
Specificity$_{95\%\ \text{CI}}$	.200	.350	$.538^{\text{sens}+}$ $.846^{\text{spec}+}$	$.720^{\text{sens}+}$ $.800^{\text{spec}+}$	$.620^{\text{sens}+}$ $.700^{\text{spec}+}$	$.333^{\text{sens}+}$ $.569^{\text{spec}+}$	.902
Overall **in**accuracy %	77.6%	62.3%	45.4%$^{\text{sens}+}$ 16.1%$^{\text{spec}+}$	26.9%$^{\text{sens}+}$ 20.5%$^{\text{spec}+}$	35.9%$^{\text{sens}+}$ 29.4%$^{\text{spec}+}$	63.0%$^{\text{sens}+}$ 42.7%$^{\text{spec}+}$	10.4%
False-negatives %	7.1%	27.5%	16.1%$^{\text{sens}+}$ 41.9%$^{\text{spec}+}$	18.0%$^{\text{sens}+}$ 24.2%$^{\text{spec}+}$	18.0%$^{\text{sens}+}$ 23.9%$^{\text{spec}+}$	14.4%$^{\text{sens}+}$ 36.5%$^{\text{spec}+}$	18.4%
False-positives %	80.0%	64.7%	46.2%$^{\text{sens}+}$ 15.4%$^{\text{spec}+}$	28.0%$^{\text{sens}+}$ 20.0%$^{\text{spec}+}$	38.0%$^{\text{sens}+}$ 30.0%$^{\text{spec}+}$	66.7%$^{\text{sens}+}$ 43.1%$^{\text{spec}+}$	9.8%

Note. N = sample size. Cut-off options: $^{\text{sens}+}$ better sensitivity and $^{\text{spec}+}$ better specificity. OR = odds ratio; 95% CI = 95% confidence interval; AUC = area under curve; ROC = receiver operating characteristic.

to adequately predict diagnostic classification. Scrutiny of the data suggests that children hit the ceiling around Year 3 on these two scales, whereas they continued to improve their scores until Year 7 on the other scales.

Discussion

Screening is an important first step in the overall workflow of supporting students along the assessment to intervention continuum. In transparent orthographies, students with dyslexia may go under the radar for a long time because they read with great accuracy from the very beginning (Caravolas et al., 2019; Torppa et al., 2016). Poor reading performance may not be readily apparent until end of elementary school, when text amount increases and reduced reading *speed* is more likely to attract attention (Reis et al., 2020). Problems with reading-related skills such as letter knowledge, phonological awareness, and rapid naming can however be seen much earlier than actual reading performance. The present test battery was therefore developed to briefly assess skills known to be proximal causes of poor reading reading and spelling, which are the cardinal features of dyslexia. The aim of this study was to develop a dyslexia marker test that can identify students who need special instructional attention. A further aim was to evaluate the new tool's psychometric properties.

Evaluation of the Psychometric Properties and Some Reflections

A standard criterion for evaluating the quality of a psychometric test is the AUC estimate, which is the area under the ROC curve (Youngstrom, 2014). Conceptually, the AUC can be interpreted as the probability that a randomly selected case with dyslexia would have either a lower or higher score on a given index measure than a randomly selected case without dyslexia. ROC curves are often used to visualize the tradeoffs between sensitivitiy and specificity in a binary classifier (i.e., between true positives versus false positives). Traditional benchmarks for gauging AUCs suggest that values $\geq .9$ are excellent, $\geq .8$ are good, $\geq .7$ are fair, and $< .7$ are poor (Youngstrom, 2014). The AUC for The 1-Minute Word Decoding Test was .922, which is "excellent" according to the traditional benchmarks (Youngstrom, 2014). This means that there is a 92.2 % probability that a randomly selected case with dyslexia is rated as "affected" by The 1-Minute Test than is a randomly selected nonaffected individual. Furthermore, the AUC estimates for the Rapid Automatized Naming Test and the Spelling Test indicate that the Dysmate-C has "good" accuracy in separating students, whereas the Phoneme Deletion Test has "fair" discrimination ability.

To further gauge the Dysmate-C's accuracy in classifying students, the sensitivity and specificity levels were evaluated. Results showed that measures of decoding and spelling and of phoneme deletion and rapid naming discriminated the true state of students with great accuracy. For example, 81.6% of the children with a formal dyslexia diagnosis (true positives) were correctly flagged, whereas 90.2% of the typical readers (true negatives) were correctly flagged by the 1-Minute Word Decoding Test. That said, due to the inverse relationship between sensitivity and specificity this test produced 18.4% false negatives and 9.8% false positives. The sensitivity should ideally be higher to find those students who need proper and timely intervention. The cut-off score should therefore be adjusted to identify as many true at-risk children as possible even if this means that students who are not at risk will be inappropriately flagged, too.

The IRT analyses proved that the individual tests measure single underlying dimensions and thus meet the basic assumption of uni-dimensionality. The different scales had good to excellent reliability with Cronbach's alphas ranging between .84 and .95, and similarly, the TICs showed excellent reliability with theta-score areas varying between −2 to 0. Moreover, an analysis of two other psychometric artifacts, namely guessing and DIF (or MIMIC), showed minor problems related to potential biases. One of the tests (i.e., PI) used a multiple-choice response format, but neither free nor fixed guessing parameters improved model fit; hence, guessing is a negligible problem for this test. As expected, gender did not explain trait score on any tests except the spelling test which favored girls. These findings support the validity of the test scores as true indicators of dyslexia problems.

Some results reported here are worth mentioning. First, the 1-Minute Word Decoding Test appeared to be the most sensitive predictor of the dichotomous group variable ("dyslexia" or "not dyslexia"). This finding adds to empirical studies from transparent European orthographies where speed problems seem to be more evident and relevant than accuracy problems (Caravolas et al., 2019; Landerl et al., 2013).

Second, the LKN and PI failed to adequately predict diagnostic classification. It should be borne in mind that although performance on these tasks are strong predictors of later reading skills in early reading development, the influence of formal reading instruction may make the same instrument highly predictive at one testing point and ineffective at another. This was demonstrated in a study by Thompson et al. (2015) where the aim was to identify a set of predictors and to estimate the individual risk for dyslexia at different developmental stages. The authors found that prediction was best for models containing several measures and that the combination of measures that interacted to predict individual risk strongly varied by age. Similarly, although PI failed to predict diagnostic classification in this study, phoneme deletion appeared to be a strong predictor. This suggests that the latter test probably provided a more age-appropriate measure of phoneme awareness than the first, to the present sample.

The third finding worth mentioning is that working memory did not predict dyslexia status. Still, this test should not be omitted from the test battery. Working memory assessments could contribute important information about children's cognitive function and achievements may provide information that prove relevant for how well a child can adjust to, or sometimes even overcome challenges presented by dyslexia (Gray et al., 2019).

Implications for Practice

Once risks are identified, progress needs to be monitored in the early stages of formal reading instruction to ensure that letter knowledge and phoneme awareness skills are acquired and if not, to provide interventions to overcome any difficulties with these foundational skills that make children vulnerable to reading and spelling difficulties (Thompson et al., 2015). Reading problems appear to be increasingly more resistant to intervention after third grade, and early risk assessments should be combined with progress monitoring of response to intervention (RTI). Miciak and Fletcher (2020) point out that when a student's RTI suggests they are not responding adequately, then a diagnostic assessment is warranted. In line with the MDM, the cornerstone of diagnostic assessment should be evaluation of the defining symptoms—based on clinical history, observations, and reliable, and validated tests—as well as careful attention to functional impairment and possible comorbid conditions (McGrath et al., 2020).

Strengths and Limitations

This study had several limitations. First, the experimental instruments could not be validated against established measures. Test protocols for the validation sample was also not available. Second, the lack of counterbalancing of the administration of subtests prohibited quantification of the variance components associated with sequence effects versus true construct variation. Future studies should control for potential impact of order effects as suggested by Kooken et al. (2017). Notwithstanding these shortcomings, the psychometric methods used here—the IRT and the MIMIC modeling in particular—provide a strong psychometric basis for the present test battery. Moreover, the automatic scoring and uploading of the data streamlined the data collection process, thus making it less error prone.

Conclusion

The Dysmate-C produces test scores that are precise reflections of the individual's latent skills within the different domains, and the test battery identifies students at risk for dyslexia with great accuracy. Since it measures the defining markers of dyslexia, it may be used both for screening and for diagnostic testing; the distinction depends on context.

Acknowledgments

The authors want to express their sincere appreciation to all students who took part in the study and to the research assistants, teachers, and school administrators who made the data collection possible.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by a grant from The Norwegian Research Council (grant no. 309374) to the first author of this paper.

Supplemental Material

Supplemental material is available on the webpage with the online version of the article.

References

Baker, F. B. (2001). *The basics of item response theory*. ERIC.

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Caravolas, M., Lervåg, A., Mikulajová, M., Defior, S., Seidlová-Málková, G., & Hulme, C. (2019). A cross-linguistic, longitudinal study of the foundations of decoding and reading comprehension ability. *Scientific Studies of Reading*, *23*(5), 386–402. https://doi.org/10.1080/10888438.2019.1580284

Catts, H., McIlraith, A., Bridges, M. S., & Nielsen, D. C. (2017). Viewing a phonological deficit within a multifactorial model of dyslexia. *Reading and Writing*, *30*(3), 613–629. https://doi.org/10.1007/s11145-016-9692-2

Georgiou, G. K., Martinez, D., Vieira, A. P. A., & Guo, K. (2021). Is orthographic knowledge a strength or a weakness in individuals with dyslexia? Evidence from a meta-analysis. *Annals of Dyslexia*, *71*(1), 5–27. https://doi.org/10.1007/s11881-021-00220-6

Gray, S., Fox Annie, B., Green†, S., Alt, M., Hogan Tiffany, P., Petscher, Y., & Cowan, N. (2019). Working memory profiles of children with dyslexia, developmental language disorder, or both. *Journal of Speech, Language, and Hearing Research*, *62*(6), 1839–1858. https://doi.org/10.1044/2019_JSLHR-L-18-0148

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47.

Kooken, J., Welsh, M. E., McCoach, D. B., Miller, F. G., Chafouleas, S. M., Riley-Tillman, T. C., & Fabiano, G. (2017). Test order in teacher-rated behavior assessments: Is counterbalancing necessary? *Psychological Assessment*, *29*(1), 98–109. https://doi.org/10.1037/pas0000314

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H. T., Lohvansuu, K., O'Donovan, M., Williams, J., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., Tóth, D., Honbolygó, F., Csépe, V., Bogliotti, C., Iannuzzi, S., Chaix, Y., Démonet, J.-F., & Schulte-Körne, G. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry*, *54*(6), 686–694. https://doi.org/10.1111/jcpp.12029

Lovett, M. W., Frijters, J. C., Wolf, M., Steinbach, K. A., Sevcik, R. A., & Morris, R. D. (2017). Early intervention for children at risk for reading disabilities: The impact of grade at intervention and individual differences on intervention outcomes.

Journal of Educational Psychology, *109*(7), 889–914. https://doi.org/10.1037/edu0000181

Magnusdottir, B. B., Haraldsson, H. M., & Sigurdsson, E. (2019). Trail making test, stroop, and verbal fluency: Regression-based norms for the Icelandic population. *Archives of Clinical Neuropsychology*, *36*(2), 253–266. https://doi.org/10.1093/arclin/acz049

Mathes, P., Denton, C., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly*, *40*(2), 148–182. https://doi.org/https://doi.org/10.1598/RRQ.40.2.2

McGrath, L. M., Peterson, R. L., & Pennington, B. F. (2020). The multiple deficit model: Progress, problems, and prospects. *Scientific Studies of Reading*, *24*(1), 7–13. https://doi.org/10.1080/10888438.2019.1706180

Miciak, J., & Fletcher, J. M. (2020). The critical role of instructional response for identifying dyslexia and other learning disabilities. *Journal of Learning Disabilities*, *53*(5), 343–353. https://doi.org/10.1177/0022219420906801

Muthén, L. K., & Muthén, B. (2017). *Mplus user's guide: Statistical analysis with latent variables, user's guide*.

Nergård-Nilssen, T., & Hulme, C. (2014). Developmental dyslexia in adults: Behavioural manifestations and cognitive correlates. *Dyslexia*, *20*(3), 191–207. https://doi.org/https://doi.org/10.1002/dys.1477

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The Patient-Patient-Centered Outcomes Research*, *7*(1), 23–35. https://doi.org/10.1007/s40271-013-0041-

Parrila, R., Dudley, D., Song, S., & Georgiou, G. K. (2020). A meta-analysis of reading-level match dyslexia studies in consistent alphabetic orthographies. *Annals of Dyslexia*, *70*(1), 1–26. https://doi.org/10.1007/s11881-019-00187-5

Peng, P., & Fuchs, D. (2014). A meta-analysis of working memory deficits in children with learning difficulties: Is there a difference between verbal domain and numerical domain? *Journal of Learning Disabilities*, *49*(1), 3–20. https://doi.org/10.1177/0022219414521667

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition*, *101*(2), 385–413. https://doi.org/10.1016/j.cognition.2006.04.008

Reis, A., Araújo, S., Morais, I. S., & Faísca, L. (2020). Reading and reading-related skills in adults with dyslexia from different orthographic systems: A review and meta-analysis. *Annals of Dyslexia*, *70*(3), 339–368. https://doi.org/10.1007/s11881-020-00205-x

Snowling, M. J., & Hulme, C. (2021). Annual research review: Reading disorders revisited—The critical importance of oral language. *Journal of Child Psychology and Psychiatry*, *62*(5), 635–653. https://doi.org/10.1111/jcpp.13324

Snowling, M. J., Hulme, C., & Nation, K. (2020). Defining and understanding dyslexia: Past, present and future. *Oxford Review of Education*, *46*(4), 501–513. https://doi.org/10.1080/03054985.2020.1765756

Snowling, M. J., & Melby-Lervåg, M. (2016). Oral language deficits in familial dyslexia: A meta-analysis and review. *Psychological Bulletin*, *142*(5), 498–545. https://doi.org/10.1037/bul0000037

Solem, C. (2021). *A report on practice for the assessment of specific reading and writing difficulties, mathematical difficulties and language difficulties*. Dysleksi Norge. https://dysleksinorge.no/wp-content/uploads/2021/03/Rapport_utredning-spraksis_2021.pdf

Thissen, D. (2000). *Reliability and measurement precision*. Lawrence Erlbaum.

Thompson, P. A., Hulme, C., Nash, H. M., Gooch, D., Hayiou-Thomas, E., & Snowling, M. J. (2015). Developmental dyslexia: Predicting individual risk. *Journal of Child Psychology and Psychiatry*, *56*(9), 976–987. https://doi.org/10.1111/jcpp.12412

Torppa, M., Georgiou, G. K., Lerkkanen, M.-K., Niemi, P., Poikkeus, A.-M., & Nurmi, J.-E. (2016). Examining the simple view of reading in a transparent orthography: A longitudinal study from kindergarten to grade 3. *Merrill-Palmer Quarterly*, *62*(2), 179–206. https://doi.org/10.13110/merrpalmquar1982.62.2.0179

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204–221.