Evaluation of the psychometric properties of "the Norwegian screening test for dyslexia"

Trude Nergård-Nilssen[1] Kenneth Eklund[2]

[1]Department of Education, UiT The Arctic University of Norway, Tromsø, Norway
[2]Department of Psychology, University of Jyväskylä, Jyväskylä, Finland

Abstract

The aim of this study was to develop and investigate the psychometric properties of a screening protocol for Norwegian students in upper secondary school. The protocol was designed to assess skills that are at stake in dyslexia. It was administered to 232 students. In the lack of a ‘gold standard’, comparisons were made between students who reported normal literacy skills (n=184) and literacy problems (n=48). Significant group differences were found across all areas. Logistic regression and ROC curve analyses demonstrated good discriminatory power. The screener meet standards for reliability and validity. It has the potential to be a useful tool for teachers to identify students at risk for dyslexia, and who thus need to be referred to full diagnostic investigation.

Keywords: screening test, dyslexia, upper secondary

Evaluation of the Psychometric Properties of 'The Norwegian Screening Test for Dyslexia'

It is important that individuals with dyslexia are identified and provided with appropriate intervention. Currently, however, many Norwegians with dyslexia are missed due to the dearth of norm-referenced screening tools. A recent study by Nergård-Nilssen and Hulme (2014) examined the nature of literacy impairments and cognitive deficits in volunteering parents in a longitudinal family risk study of dyslexia. Only three out of the 55 parents who claimed to have dyslexia, and whom fulfilled the diagnostic criteria, had a formal diagnosis before they were enrolled in the family study. This clearly demonstrates the ignorance of literacy disorders and need for satisfactory screening and diagnostic tools. The purpose of this study was to develop and collect initial reliability and validity data on an instrument designed to screen for dyslexia in Norwegian students.

The construction of the protocol was guided by theories that suggest that dyslexia is a heritable disorder with a multifactorial aetiology, and is the outcome of multiple risk factors (Catts, McIlraith, Bridges, & Nielsen, 2017; Hulme & Snowling, 2009; Thompson et al., 2015; van Bergen, van der Leij, & de Jong, 2014). The present protocol complies with the current definitions of developmental dyslexia in the U.S. and in UK, respectively. The two working definitions are very similar and build on accumulated research. Both definitions state that dyslexia is a learning difficulty that primarily affects the skills involved in accurate and fluent word reading and spelling, and that characteristic features include difficulties in phonological processing (Panel, Health, & Development, 2000; Rose, 2009). The skills at stake in dyslexia are continuous and dimensional in nature, and it is thus possible to establish how far and in what direction (positive vs. negative) a measured value deviates from the population mean. The outcome profile depends upon the status of the phonological and broader oral language skills that the individual brings to the task of reading (Catts, Adlof, & Weismer, 2006; Hulme & Snowling, 2014; Snowling & Hulme, 2012), and each individual

has a unique profile of strengths and weaknesses. Although the continuum of performance means that any cut-off point is arbitrary (Siegel, 2006), we expect that affected individuals in general would achieve scores in the lower end of the normal distribution on tests measuring decoding, spelling and phonological processing.

The selection of tests for the present protocol was based on the findings from the Nergård-Nilssen and Hulme study (2014). Among other things, Nergård-Nilssen and Hulme found that spelling skills distinguished the dyslexic parents markedly from the typical parents (d = 1.82), suggesting that spelling disorders are the most prominent residual marker of dyslexia even in compensated Norwegian adults. A similar word-spelling task was therefore developed for the present protocol to assess students' explicit knowledge of the orthographic structure of high-frequency exception and regular words. Further, the Nergård-Nilssen and Hulme study demonstrated that problems with phoneme awareness are persistent correlates of literacy disorders in adults. For the present protocol, a pseudohomophone test was developed to measure phonological processing by means of silent nonword decoding. Further, a writing efficiency test was designed to assess how well word spellings are automatized and how well students write under time constraints. The rationale behind this test was the finding by Warmington, Stothard, and Snowling (2013) that university students in the UK with dyslexia performed significantly worse than students without dyslexia on writing speed (d = 1.06). Warmington et al. (2013) highlight the importance of assessing skills that are generally used in the learning environment. Finally, a reading comprehension test was developed. Although developmental dyslexia is acknowledged as an impairment that affects the development of accurate and fluent word level reading and spelling skills, dyslexic individuals may also have trouble with understanding what they read. Reading comprehension impairments is however generally considered a consequence of the difficulties in word recognition and the lack of reading experience in dyslexic readers (Catts, Kamhi, & Adlof, 2012; Hulme & Snowling,

2014; Snowling & Hulme, 2012). Notwithstanding this, students who perform poorly on reading comprehension need to undergo a thorough assessment that addresses the linguistic and cognitive skills that might underpin their poor outcomes. The present reading comprehension test was designed to be syntactically complex and to reflect the degree of difficulty in textbooks that are typical for this educational level, and yet place relatively modest demands on the students' decoding skills.

Three requirements guided the development of tests in the new screening protocol. Firstly, the tests should be simple enough in format to enable professionals with no prior training in testing to administer and score the results. Secondly, the tasks should yield a broad picture of literacy skills and assess abilities that are required by students who wish to pursue upper secondary school. The screening protocol reported here includes tests that measure phonological processing, spelling, decoding, writing efficiency, reading comprehension and vocabulary. Poor skills in these areas are associated with dyslexia. Finally, the students should carry out the tasks without vocal responses (i.e., silently). This would allow the test to be administered to all students in a class simultaneously. Written responses would save time during testing because the examiner would not need to register individual responses during testing but instead evaluate the responses afterwards. In addition, this would enable the examiner to introduce the tasks and practice items and to provide students with group-wise corrective feedback. A shortcoming with group-administered tests, though, is that they preclude assessments that require oral responses such as single word decoding and rapid automatized naming (RAN). Poor skills in word decoding and RAN are long term predictors of reading difficulties in Norwegian children (Furnes & Samuelsson, 2010; Lervåg, Bråten, & Hulme, 2009) and residual markers of dyslexia in Norwegian adults (Nergård-Nilssen & Hulme, 2014). At-risk students identified by the present protocol should thus be tested in these areas in the subsequent individual assessment.

In sum, the test protocol originally consisted of four tests including spelling, nonword decoding, writing efficiency and reading comprehension. Test items were deliberately selected to capture key aspects of the Norwegian phonology. For example, several consonants in Norwegian occur as so-called silent letters, which means they are written, but not pronounced (e.g., hva (what) /va:/; hjem (home) /jem/; huset (the house) /hu:se/). Another example is the 'phoneme length attribute', which implies that semantic content are signalled by difference in vowel length in speech (e.g., 'mine' (my; mine) /'mi:ňe/ - 'minne' (memory; remind) /'miň:e/). In Norwegian orthography, this vowel length difference is signalled by the subsequent consonants, that is, long vowel pronunciations are followed by a single consonant whereas short vowel pronunciations are followed by two identical consonants, e.g., lege (doctor) – legge (put; place). The screening protocol addresses these and other properties of the Norwegian orthography in various ways. ~~For example, the spelling test required the participants to spell four homophone words: 'hvert' (each, every), 'vert' (host), 'vært' (been), and 'verdt' (worth) – all pronounced [væt]. Here, all words were framed within a sentence to ensure the correct meaning.~~

The first wave of data collection showed a clear need for addressing decoding skills and general language in more detail. A word-chain test was thus developed to assess students' skills in word decoding and word recognition, and to assess their orthographic knowledge. Word recognition tasks are typically carried out in a one-to-one setting where students are asked to read aloud lists of words as quickly and correctly as possible within time restrictions (e.g., one minute). A word-chain test, on the other hand, does not require reading aloud. Instead, this test requires the students to separate words by drawing lines where space is removed and can thus be administered group wise. In addition, a multiple-choice vocabulary test was included to obtain information about general language skills. It is well established that vocabulary size, depth of vocabulary knowledge, and reading comprehension are highly

and positively correlated (Qian, 1999). An extended normative study was carried out to evaluate the reliability and validity after the two new tests were added to the protocol.

Evaluation of a test is typically a matter of determining the test's diagnostic accuracy, that is, its ability to discriminate between individuals that have and those who do not have the condition of interest. The best available method to determine the presence or absence of a condition is by using a reference standard, or so-called 'gold standard'. Ideally, the psychometric properties of any novel test is established by comparing how the results of the new test agree with the "true" outcome. However, as Rutjes, Reitsma, Coomarasamy, Khan, and Bossuyt (2007) point out, researchers often encounter situations where the reference standard (e.g. a diagnostic test) is not available in all patients, where the reference standard is imperfect or where there is no accepted reference standard. In the present study, the fraction of students verified with the reference standard was very small, and we did not have access to the test protocols to students who were diagnosed with literacy disorders (e.g. dyslexia). We thus faced what Rutjes et al. (2007) refer to as 'a no gold standard situation' in the evaluation of the new screening device.

In the absence of a gold standard, we examined how the test results related to self-reported and self-conceived literacy problems. That is, in this initial attempt to examine the protocol's construct validity and discriminant power, we compared two samples: One sample who reported a history of reading and/or spelling disorders ('Impaired group') and one sample who reported no problems with reading and spelling ('Non-Impaired group'). We anticipated that self-reported literacy problems should differentiate students with and without dyslexia-associated traits, and that student with self-reported literacy problems should gain lower scores on measures of reading, spelling and phonological processing.

Method

Participants

A random sample of five different upper secondary schools in the XXX County were invited to take part in the present study. In each school, all three year levels were invited to take part (i.e., Year 11-Year 13). Participants were either attending Vocational Education and Training or the Programme for Specialization in General Studies. In addition, a group of first year university students was invited to take part in this study. These students had attended the programme for specialization in general studies in upper secondary, and none of them had yet sit for examination at university level. An information sheet and a letter of consent for students younger than 18 years were distributed to parents. Two hundred and thirty two students volunteered for the study. Among these were three non-native speakers of Norwegian (these were exchange students attending the program for general studies, and reported no history of literacy disorders). The mean age for participants was 18.68 years (*SD* = 3.61 years). Before the protocol was administered, all students were asked to fill in a form enquiring family history of language impairments and dyslexia. Responses to 'Do you have current or past problems with reading, spelling and/or reading comprehension?' were dichotomized afterwards for making classification, and to predict literacy status. Table 1 shows characteristics for the two groups.

Non-Impaired group. Participants who reported no history of literacy disorders were designated to the 'Non-Impaired group'. One hundred and eighty four students met inclusion criteria for this group (96 females, 88 males).

Impaired group. Participants were designated to the 'Impaired group' if they reported current or earlier problems with either reading, writing and/or read reading comprehension. Forty-eight students met inclusion criteria for this group (30 females, 18 males). Twelve members in this group were diagnosed with dyslexia and had been deliberately invited to

participate through the local dyslexia association. These participants were students in other schools than those schools randomly selected for this study. Unfortunately, we do not have access to their test protocols or test reports.

Materials

Spelling test. The spelling-to-dictation task consist of 45 common words, which varies in terms of length and complexity. To address orthographic knowledge, the spelling test includes words with silent letters and homophone words – as for instance 'hvert' (each, every), 'vert' (host), 'vært' (been), and 'verdt' (worth), where all four words are pronounced [væt]. All words are framed within a sentence to ensure the correct meaning, e.g., *"a person who receives or entertains other people as guests is a host. Please write host"*. There are no time limits, and if necessary, the sentence and target word can be repeated once. The score is the number of correctly spelled words, with a maximum score of 45.

Word-Chain test. In this test, participants are instructed to identify and disentangle words that are written together (i.e. without space) by drawing lines between the words. Each item, or chain, consists of four high-frequency words that represent different word classes (e.g., nouns, verbs, adjectives, adverbs, prepositions or numerals) and vary in length from one to six letters (e.g**.,** *kropp|saks|rå|stilk*). Three practice items introduce the test. The students have a limit of 4 minutes to solve as many word-chains as possible. The score is the number of correctly solved chains, with a maximum score of 56.

Pseudohomophone test. This test consists of 25 tasks. In each task, participants are presented with five nonwords of which one is phonetically identical to a real word (e.g., nale, keap, gaim). The four alternative nonwords do not sound like words but are, in varying degree, visually similar to real words. Three practice items introduce the test. Students are instructed to identify as many pseudohomophones as possible within the time limit of 2

minutes. The score here is the number of correctly identified pseudohomophones, with a maximum score of 25.

Writing efficiency test. Here, students are presented with a 17-word sentence and are asked to write up the sentence as many times they can manage for two minutes, as quickly and accurately as possible. The sentence is typed at the top of the work sheet, and participants are allowed to refer back to the sentence as many times as needed. The score here is the number of correctly spelled words per minute.

Reading comprehension test. In this test, students are presented with a 1071-word text followed by 14 multiple-choice questions. Four options follow each question – where one is correct and three seem possible but are incorrect in some way (distracters). For the present purpose, a text about the early pioneer of antiseptic procedures, Ignaz Semmelweis, was developed. The multiple-choice tasks were developed to include a mix of literal, reorganizational and inferential questions as proposed by Day and Jeong-suk (2005). Questions of literal comprehension can be answered directly and explicitly from the text (e.g., "What is puerperal fever?"). Questions of organizational comprehension require students to use information from various parts of the text and combine them for additional understanding (e.g., "Which of the following hypotheses did Semmelweis not test?"). Questions of inference comprehension required participants to combine their literal understanding of the text with their own knowledge and intuitions (e.g., "Why were Semmelweis' ideas rejected by the medical community?"). Participants are instructed to read silently the text as quickly and accurately as possible, and to complete the following multiple-choice questions within the time limit of 10 minutes. Students are allowed to refer back to the text as many times as needed. The score here is the number of correct answers, with a maximum score of 14.

Vocabulary test. The multiple-choice vocabulary test consists of 15 tasks. Each task contains one stem-word (e.g., 'implement') followed by four alternative words (e.g., verify,

effectuate, blend, illustrate). Participants are asked to find and mark the synonym to each stem-word. There are no time limits for this test. The score here is the number of correctly identified synonyms, with a maximum score of 15.

Procedure

The original protocol (consisting of four tests) was administered to 111 students. In this first wave, 95 students met criteria for inclusion in the "Non-Impaired group, whereas 16 students met criteria for inclusion in the "Impaired group". After the inclusion of the two new tests, we carried out an extended normative study. In this second wave, another 121 students were included. Here, 89 students met criteria for inclusion in the "Non-Impaired" group, whereas 32 students met criteria for inclusion in the "Impaired group". In short, 232 participants carried the original four tests, whereas 121 participants carried out the revised protocol where the two new tests were added.

All tasks were group administered in a single session lasting approximately 30-40 minutes. Three research assistants, who had received extensive training for the tests being used, administered the test protocol in a fixed order. All tests were scored independently and verified by the two other research assistants.

Results

The evaluation of the psychometric properties of the screening protocol was carried out by examining its reliability, construct validity and discriminant power.

Descriptive statistics and reliability

Table 2 reports the descriptive statistics and reliability coefficients for the screening protocol. An initial normality check showed that the scores were normally distributed on all tests except the Spelling Test, which was left skewed (-1.32). Scrutiny of the data showed that four students who self-reported literacy problems and one student who reported normal skills obtained scores more than 2.5 standard deviations below the mean on this test. Since the

scores were generally normally distributed, we transformed the distribution for each test into z-scores.

Further, we estimated internal consistency to check if the items that make up the individual tests (scales) 'hang together' and measure the same underlying construct. We estimated the reliability with Cronbach's coefficient alpha. On the Reading Comprehension Test, however, questions varied in difficulty, and reliability was therefore calculated by using Guttmann's Split-half coefficient. Estimates showed that all tests have reliability coefficients between .84-.92, suggesting very good to excellent reliability. The screening protocol thus fulfils psychometric criteria for reliability.

Validity

To examine the screening protocol's construct validity and discriminant power, we conducted three kinds of analyses. First, we ran correlational analysis to assess the internal relationship between the measures in the protocol. Next, we conducted a series of independent samples *t*-tests with Cohen's *d* to examine the magnitude of any group differences. Finally, we evaluated the protocol's predictive validity by running logistic regression analysis and ROC analysis.

Inter-correlations between tests. We ran correlations to assess the concurrent relationship between the different measures in the protocol. As Table 3 shows, all correlations were significant. In general, moderate positive correlations were found between tests measuring spelling and decoding, and between reading comprehension and vocabulary. As expected, phonological processing were more strongly related with spelling and word recognition (as indexed by word chains) than any of the other measures. The correlation coefficients provides not only a measure of the relationship between the tests but also an index of the proportion of variance shared between the different tests. By squaring the correlation coefficients, it turned out e.g. that the spelling test shared between 27-35 %

variance with the other tests (but only 22 % with reading comprehension), and that vocabulary and reading comprehension share 37 % variation. In summary, the six tests showed moderate relationships with each other but no signs of multicollinearity.

Independent samples t-tests. To examine group differences, we ran independent samples *t*-tests and computed Cohen's *d* for effect size. Table 4 shows the performance of the two groups and the standardized group differences. As hypothesized, students who reported literacy problems performed significantly less well on all six tests compared to students who self-reported normal skills. Effect sizes were large for all group comparisons, ranging from 0.94 to 1.61. The large effect sizes clearly indicates that the distribution of scores for the two groups were different and that the six tests clearly distinguish the two groups.

Logistic regression and ROC curve analysis. We conducted logistic regression and ROC curve analyses to further evaluate the accuracy of the present protocol. Generally, a logistic regression model calculates the group membership probability and provides an estimate of accuracy for decision-making. Accuracy is evaluated by the model fit, as well as indices of sensitivity and specificity. A ROC curve model, on the other hand, plays a central role in evaluating the ability of tests to discriminate the true state of subjects. A ROC curve model finds the optimal cut-off values and provides a combined measure of sensitivity and specificity that describes the inherent validity of the screening instrument (Hajian-Tilaki, 2013). As such, some would argue that the ROC curve analysis is more informative than the logistic regression classification table since it summarizes the predictive power for all classification probability values.

Given the lack of a 'gold standard', we used the dichotomous variable "Impaired" and "Non-Impaired" according to the student's response to the question "Do you have current or past problems with reading, spelling and/or reading comprehension?" The model contained six independent variables, that is, the six tests included in the screening protocol. Here, only

participants who had completed the full screening protocol were entered into the analysis (N=121). The full model was statistically significant, χ^2 (6, N=121) = 67.22, p < .001, indicating that it was able to distinguish between students who reported literacy problems from those who reported normal literacy skills. The model as a whole explained between 42.6 % (Cox and Snell R^2) and 62.2 % (Nagelkerke R^2) of the variance in self-reported literacy status.

With a set cut-off value of .50, the model correctly classified 84.3 % of cases overall, with a sensitivity level of 62.5 % and a specificity level of 92.1 %. However, Jenkins and colleagues recommend that a minimum sensitivity should be as high as 90 percent in screening for literacy disorders, when the purpose is to ensure that truly at-risk students are identified (Jenkins, Hudson, & Johnson, 2007; Johnson, Jenkins, Petscher, & Catts, 2009). We therefore selected a cut-off score of .12 to increase the sensitivity level. The increase in sensitivity level to 90.6 % led to a decrease in the specificity level to 70 % (and an overall accuracy to 75.2 %). Scrutiny of the data showed, however, that some of the misclassified students – who self-reported normal skills – turned out to have poor skills according to the screening protocol. Consequently, we cannot leave out the possibility that the misclassified students in fact do have literacy disorders, and that the specificity measure reported here is somewhat flawed.

Next, we conducted a ROC curve analysis to further evaluate the protocol's accuracy. This analysis estimates that a randomly chosen member of one group has a higher probability of belonging to that group than has a randomly chosen member of the other group. The accuracy is indexed by the area under the ROC curve. An area of 1.00 represents a perfect discrimination whereas an area of .50 represents no discrimination. Results showed that the area under the curve (AUC) was .92 for the full model (with 95% CI from .87 to .97, SE = .025, p <.001). In other words, there is a 92 % probability that a randomly chosen affected

individual is rated as more likely to be affected than a randomly chosen non-affected individual is by the present protocol. An AUC of .92 is excellent according to guidelines for classifying the accuracy of diagnostic tests (Tape, 2006).

To judge whether any single test would separate the two groups nearly as accurately as the entire protocol, we then ran a ROC analysis to identify the AUC for each test. An attractive advantage of ROC curve analysis is that one can compare individual tests and assess whether the various combination of tests can improve diagnostic accuracy when each test is performed on the same subjects (Hajian-Tilaki, 2013). Table 5 shows the AUC for each of the tests. As can be seen, word decoding, spelling and phonological processing had high AUCs (ranging from .849 to .875) while the other three tests only had fair AUCs (ranging from .734 to .753). We then compared the AUCs of each dependent test against that of each other and against the entire protocol by means of the MedCalc statistical package, which offers a nonparametric comparison between ROC curves based on the method developed by Hanley and McNeil (1982) and DeLong, DeLong, and Clarke-Pearson (1988). The statistical significance of the difference between ROC curves was calculated with the *z* test. The fifteen possible pairwise comparisons between tests showed that the ROC curves for word decoding, spelling and phonological processing, respectively, were not significantly different from each another, with *z* scores ranging from 0.24 to 0.64. Similarly, the ROC curves for the three remaining tests, writing efficiency, reading comprehension and vocabulary, were not significantly different from one another, with *z* scores ranging from 0.19 to 0.32. The ROC curves for spelling and word decoding, but not phonological processing, were however significantly different from all the three latter tests, with *z* scores ranging from 2.05 to 2.94. In a last step, we compared each test against the entire protocol. Table 5 shows the outcomes of the six pairwise ROC curve comparisons against the protocol. As can be seen, all compared areas are significantly different. The results clearly show that the ROC curve area for the

protocol is significantly greater than the area for any individual test and, consequently, that no test should be omitted from the protocol.

Discussion

To this date many Norwegians with dyslexia are missed due to the lack of agreed upon procedures and lack of norm-referenced screening tools. The need to identify students in upper secondary (high school) as having literacy difficulties is concerning. Students should ideally be identified and receiving special education assistance much earlier. Nonetheless, the main aim of the present study was to develop and evaluate the psychometric properties of a screening protocol for students in upper secondary (high school). Due to the lack of a gold standard, we could not validate outcomes on the protocol against a criterion-validated test. Instead, we focused on how reliably each test measured attributes that are associated with dyslexia and how well outcomes would reflect our *a priori* expectations.

As expected, students who reported literacy problems performed significantly less well on all six tests compared to students who self-reported normal skills. Effect sizes were large for all group comparisons. Results further showed that phonological processing was more strongly related to spelling and word recognition than to any other measure. This is not surprising given the accumulated empirical findings that phonological processing abilities exert strong causal influences on word decoding and spelling (Caravolas, Hulme, & Snowling, 2001; Melby-Lervåg, Lyster, & Hulme, 2012; Wagner, Torgesen, & Rashotte, 1994). We also found that decoding, spelling and phonological processing discriminated the two groups better than writing efficiency, reading comprehension and vocabulary. Again, this was not surprising since it is well established that these are core markers of dyslexia at the behavioural and cognitive level, respectively. Noteworthy is however the observation that self-reported poor readers performed significantly less well on vocabulary. This test has no time limits, so a more likely explanation might be that their limited vocabulary is a

consequence of prolonged reading problems. That is, these students are less exposed to printed words and thus words encountered in contexts that expose word meanings (Ocal & Ehri, 2017). Alternatively, the poor group performance may reflect a co-occurrence of oral language and reading difficulties in students (Catts, Adlof, Hogan, & Weismer, 2005). Notwithstanding, students who obtain scores more than one standard deviation below the mean on the vocabulary test or any other test should be assessed in greater depth in an individual follow-up.

The screening protocol proved to meet standards for reliability, with coefficients ranging from .84 to .92. Similarly, the protocol proved to meet standards for validity. As Cronbach and Meehl (1955) highlight, however, construct validity cannot generally be expressed in the form of a single simple coefficient (although a numerical estimate can sometimes be arrived at by a factor analysis). Instead, many types of evidence are relevant to construct validity, including inter-item correlations, inter-test correlations, test-"criterion" correlations, and group comparisons. Cronbach and Meehl (1955) point out that a construct is some postulated attribute of people assumed to be reflected in test performance, and that in test validation, the attribute about which we make statements in interpreting a test is a construct. As such, it is naïve to inquire, "Is this test valid?" because one does not validate a test, but only the principle for making inferences (Cronbach & Meehl, 1955; Kane, 2013). What is to be validated in the present case is the inferences and interpretations derived from the test outcomes: that is, inference of whether a student is at risk for dyslexia or not, and consequently, whether this student should be referred to a full assessment or not.

Convergent evidence for validity was derived from an evaluation of the accuracy with which the screening protocol discriminates between students with and without literacy disorders. Here, independent samples *t*-tests showed that the distribution of scores were significantly different, and that effect sizes were large for all group comparisons. Further, the

logistic regression analysis demonstrated good model fit and that the protocol predicted group membership with satisfactory accuracy. Similarly, ROC curve analysis showed that the protocol discriminated the true state of subjects with great accuracy, and that the ROC curves had excellent AUC. Divergent evidence for validity was derived from inter-test correlations and from comparisons between the AUCs and ROC curves for tests. The comparisons clearly show that the overall performance of the protocol is significantly better than any individual test. Taken together, the empirical data indicates that the protocol identifies students at risk for dyslexia with high accuracy and that validity thus is good.

It is interesting to speculate why poor achieving students, who self-reported normal reading and writing skills, had not recognised their problems before they took part in the present study. There is no simple answer to this. However, it is important to note here that four out of six tests in the protocol were time-limited. Mounting studies suggest that reading speed poses a greater challenge than reading accuracy in transparent orthographies for both normal readers (Seymour, Aro, & Erskine, 2003) and dyslexic readers (Ziegler, Perry, Ma-Wyatt, Ladner, & Schulte-Korne, 2003). It is likely that the semi-transparent nature of the Norwegian orthography makes students overestimate their reading accuracy, and thus their general literacy skills.

Clearly, the reading and writing problems observed in the present sample had slipped below the radar during primary school and lower secondary school. The lack of clarification of how developmental dyslexia should properly be assessed, the lack of agreed procedures for identifying dyslexia and reading comprehension impairments, respectively, and the dearth of norm-referenced screening tools in Norwegian is striking. As a rule, the school refers a student to the local Educational-Psychological Service when dyslexia is suspected. The educational psychologist then normally carries out a standardised diagnostic test, (i.e., Logos developed by Høien, 2007). This test requires a one-to-one administration procedure.

However, individual-administered tests do not necessarily have better psychometrical qualities, nor higher construct validity, than group-administered tests (Belmont & Borkowski, 1988). Group-administered tests are thus an attractive alternative (Wolff & Lundberg, 2003). The main aim with the present screening protocol is to provide a quick and inexpensive measure for identifying risk markers of dyslexia, which can be carried out by teachers.

Like any other screening instrument, the present protocol suffers from limitations due to the inverse relationship between sensitivity and specificity (Bogue, 2011; Jenkins et al., 2007; Johnson et al., 2009; Plante & Vance, 1994). A screening instrument with high sensitivity levels yield false positives. This wastes resources by providing them to individuals who do not need them. In contrast, a screening instrument with high specificity yield too many false negatives. This denies assistance to those who in fact need it. Given the current situation in Norway, a screening instrument with high sensitivity is preferable. Arguably, a screening test should only be regarded a first stage of the diagnostic process. False positives should be reported off the list in the subsequent follow-up. It is crucial not to miss students who in fact have dyslexia.

The major weakness with the present study is however the lack of a reference standard (or so-called ‘gold standard’) against which to validate test outcomes. The grouping in this study is based on the students’ subjective judgements of their own literacy skills. A serious objection can be raised to this: What actually is the practical value of this screening protocol if what it does is to distinguish people who say they have literacy problems from those who do not? Why don't we just ask them, which could be done in less than a minute? There are three counterclaims to this. First, due to the earlier described circumstances, we were forced to base the grouping on student’s subjective statements. However, like the subjective feeling of chest pain is a ‘warning sign’ of myocardial infarction, although not sufficient to define the presence of the disease, students’ report of personally perceived and experienced problems

with reading or writing might be a ‘warning sign’ of dyslexia. Indeed, our data showed that the two groups differed significantly on all tests. Second, for too long have teachers in Norway identified students with dyslexia by chance, depending on the teachers’ knowledge about dyslexia. By screening the entire class with a norm-referenced protocol, students at risk will be identified independent of the teacher’s personal attitude, beliefs and knowledge. Finally yet importantly, the finding that misclassified students – who self-reported normal skills but turned out to have poor skills – clearly demonstrates that it is not sufficient to ask students if they think they have dyslexia. The general ignorance of dyslexia in Norway may influence how students respond to this question.

Dyslexia does not resolve but is persistent into adulthood (Bruck, 1990, 1992; Kemp, Parrila, & Kirby, 2009; Nergard-Nilssen & Hulme, 2014). An increasing number of Norwegian students with undetected dyslexia enter higher education institutions. The ability to detect dyslexia is important to educational institutions and yet few, if any, standardised screeners for adults are available. This problem appears to apply to countries other than Norway too (Fernandes, Araújo, Sucena, Reis, & Castro, 2017; Reynolds & Caravolas, 2016; Warmington et al., 2013). In future, the present protocol will be adapted for use in higher education. The inclusion of a bigger sample will furthermore enable construct validity investigations by utilizing confirmatory and exploratory factor analysis. These statistical procedures are of particular interest for the purposes of psychometric instruments and clinical measures, and results will be included in the test manual.

In conclusion, based on the empirical evidence, the present protocol meets the standards for reliability and validity. In addition, it is easy to administer and is time and cost effective. It has the potential to be a useful and valid tool for identifying Norwegian students with undetected literacy disorders, and who thus need to be referred to a full assessment by an expert.

Acknowledgements

This study was supported by a grant from the xx Research Foundation (grant number A42966) and NTTAS (grant number 309 FNY-3405- Verifikasjonsstudie) to the first author of the paper. We want to express our sincere appreciation to all students who took part in the study and to the teachers and school administrator who made the data collection possible. We would also like to express our gratitude for all the valuable and constructive comments we have received from HWC on drafts of this paper. His comments have greatly improved the manuscript. Further, we thank IMK, ENN, SLB and KS for assistance with data collection, data managing and coding.

References

Belmont, J., & Borkowski, J. (1988). A group-administered test of children's metamemory. *Bulletin of the Psychonomic Society, 26*(3), 206-208. doi:10.3758/BF03337288

Bogue, E. (2011). A psychometric analysis of childhood vocabulary tests.

Bruck, M. (1990). Word-recognition skills of adults with childhood diagnoses of dyslexia. *Developmental Psychology, 26*(3), 439-454. doi:10.1037/0012-1649.26.3.439

Bruck, M. (1992). Persistence of dyslexics' phonological awareness deficits. *Developmental Psychology, 28*(5), 874-886. doi:10.1037/0012-1649.28.5.874

Caravolas, M., Hulme, C., & Snowling, M. J. (2001). The Foundations of Spelling Ability: Evidence from a 3-Year Longitudinal Study. *Journal of Memory and Language, 45*(4), 751-774. doi:http://dx.doi.org/10.1006/jmla.2000.2785

Catts, H. W., Adlof, S. M., Hogan, T. P., & Weismer, S. E. (2005). Are Specific Language Impairment and Dyslexia Distinct Disorders? *Journal of Speech, Language, and Hearing Research, 48*(6), 1378-1396. doi:10.1044/1092-4388(2005/096)

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech Language and Hearing Research, 49*(2), 278-293. doi:10.1044/1092-4388(2006/023)

Catts, H. W., Kamhi, A. G., & Adlof, S. (2012). *Defining and classifying reading disabilities* (A. G. Kamhi & H. W. Catts Eds. 3rd ed.): Allyn & Bacon Communication Sciences and Disorders.

Catts, H. W., 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. doi:10.1007/s11145-016-9692-2

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281.

Day, R., & Jeong-suk, P. (2005). Developing reading comprehension questions. *Reading in a foreign language, 17*(1), 60.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

Fernandes, T., Araújo, S., Sucena, A., Reis, A., & Castro, S. L. (2017). The 1-min Screening Test for Reading Problems in College Students: Psychometric Properties of the 1-min TIL. *Dyslexia*.

Furnes, B., & Samuelsson, S. (2010). Predicting reading and spelling difficulties in transparent and opaque orthographies: a comparison between Scandinavian and US/Australian children. *Dyslexia, 16*(2), 119-142. doi:10.1002/dys.401

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine, 4*(2), 627.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Hulme, C., & Snowling, M. J. (2009). *Developmental disorders of language, learning and cognition*. Oxford: Wiley-Blackwell.

Hulme, C., & Snowling, M. J. (2014). The interface between spoken and written language: developmental disorders. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 369*(1634), 20120395.

Høien, T. (2007). *Logos håndbok: Diagnostisering av dysleksi og andre lesevansker.* . Bryne:: Logometrica AS.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*(4), 582.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice, 24*(4), 174-185.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kemp, N., Parrila, R. K., & Kirby, J. R. (2009). Phonological and orthographic spelling in high-functioning adult dyslexics. *Dyslexia, 15*(2), 105-128. doi:10.1002/dys.364

Lervåg, A., Bråten, I., & Hulme, C. (2009). The cognitive and linguistic foundations of early reading development: A Norwegian latent variable longitudinal study. *Developmental Psychology, 45*(3), 764-781. doi:10.1037/a0014132

Melby-Lervåg, M., Lyster, S.-A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: a meta-analytic review: American Psychological Association.

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

Ocal, T., & Ehri, L. (2017). Spelling Ability in College Students Predicted by Decoding, Print Exposure, and Vocabulary. *Journal of College Reading and Learning, 47*(1), 58-74. doi:10.1080/10790195.2016.1219242

Panel, N. R., Health, N. I. o. C., & Development, H. (2000). *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*: National Institute of Child Health and Human Development, National Institutes of Health.

Plante, E., & Vance, R. (1994). Selection of Preschool Language TestsA Data-Based Approach. *Language, Speech, and Hearing Services in Schools, 25*(1), 15-24.

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56*(2), 282-308.

Reynolds, A. E., & Caravolas, M. (2016). Evaluation of the Bangor Dyslexia Test (BDT) for use with Adults. *Dyslexia, 22*(1), 27-46. doi:10.1002/dys.1520

Rose, J. (2009). Identifying and teaching children and young people with dyslexia and literacy difficulties: an independent report.

Rutjes, A., Reitsma, J., Coomarasamy, A., Khan, K., & Bossuyt, P. (2007). Evaluation of diagnostic tests when there is no gold standard. A review of methods. *HEALTH TECHNOLOGY ASSESSMENT-SOUTHAMPTON-, 11*(50).

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*(2), 143-174. doi:10.1348/000712603321661859

Siegel, L. S. (2006). Perspectives on dyslexia. *Paediatrics & child health, 11*(9), 581.

Snowling, M. J., & Hulme, C. (2012). Annual Research Review: The nature and classification of reading disorders – a commentary on proposals for DSM-5. *Journal of Child Psychology and Psychiatry, 53*(5), 593-607. doi:10.1111/j.1469-7610.2011.02495.x

Tape, T. G. (2006). Interpreting diagnostic tests. *University of Nebraska Medical Center, http://gim. unmc. edu/dxtests*.

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. doi:10.1111/jcpp.12412

van Bergen, E., van der Leij, A., & de Jong, P. F. (2014). The intergenerational multiple deficit model and the case of dyslexia. *Frontiers in Human Neuroscience, 8*, 346. doi:10.3389/fnhum.2014.00346

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental Psychology, 30*(1), 73.

Warmington, M., Stothard, S. E., & Snowling, M. J. (2013). Assessing dyslexia in higher education: the York adult assessment battery-revised. *Journal of Research in Special Educational Needs, 13*(1), 48-56. doi:10.1111/j.1471-3802.2012.01264.x

Wolff, U., & Lundberg, I. (2003). A technique for group screening of dyslexia among adults. *Annals of Dyslexia, 53*(1), 324-339. doi:10.1007/s11881-003-0015-3

Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Korne, G. (2003). Developmental dyslexia in different languages: language-specific or universal? *J Exp Child Psychol, 86*(3), 169-193.

Table 1.

Characteristics of participants

		Group			
Characteristic		Non-impaired (n = 184)		Impaired (n =48)	
		Female	Male	Female	Male
Gender	n	96	88	30	18
	%	52.2	47.8	62.5	37.5
		Vocational	Academic	Vocational	Academic
Education	n	23	161	26	22
	%	12.5	87.5	54.2	45.8
		Mean	SD	Mean	SD
Age (years)		18.37	3.82	17.70	2.56

Table 2.

Descriptive statistics and reliability coefficients for the tests.

Test	Mean (Median)	SD	Min-Max	Skewness	Kurtosis	Reliability
Spelling	35.90 (38)	6.42	13-44	-1.32	1.72	.87
Word-Chain	20.04 (21)	6.92	0-33	-.62	.27	.90
Pseudohomophones	12.60 (13)	4.83	2-24	-.13	-.58	.91
Writing Efficiency	18.92 (19)	4.53	7-29	-.38	.44	
Reading Comprehension	8.65 (9)	3.40	0-14	-.87	.35	.92
Vocabulary	9.79 (10)	3.14	0-15	-.66	.48	.84

Table 3.

Correlations between tests included in the screening battery

	2.	3.	4.	5.	6.
1. Spelling	.59***	.52***	.47***	.57***	.53***
2. Word-Chain	-	.58***	.58***	.59***	.44**
3. Pseudohomophone detection		-	.35**	.45**	.31**
4. Writing efficiency			-	.34**	.49***
5. Reading comprehension				-	.61***
6. Vocabulary					-

Note. Number of participants is 232 in all correlations except Word recognition and Vocabulary where number of participants is 121.

** p <.01, *** p < .001

Table 4.

Descriptive statistics and group comparisons with independent-samples t-test and effect size (Cohen's d)

	Group						
	Non-Impaired		Impaired				Effect size
	Mean	*SD*	*Mean*	*SD*	*t*	*df*	(Cohen's *d*)
Spelling	38.29	4.61	29.75	6.70	10.20***	230	1.47
Word-Chain	22.54	5.13	13.09	6.55	8.28***	119	1.61
Pseudohomophone Detection	13.78	5.11	7.54	3.71	7.93***	230	1.40
Writing Efficiency	20.41	3.74	15.68	5.01	7.20***	230	1.07

Reading Comprehension	9.28	2.52	5.63	3.83	7.96***	230	1.13
Vocabulary	10.56	2.55	7.63	3.62	4.97***	119	0.94

Note. N = 232 (Non-Impaired, n=184 and Impaired, n=48) except for Word-Chain Test and Vocabulary where N = 121 (Non-Impaired, n=89 and Impaired, n=32).

*** p < .001

Table 5.

Area under the curve (AUC) for the six tests in the screening protocol.

Test	Area	*Asymp. Sig.*	95 % CI	S.E.
Spelling	.860	< .001	.785-.916	.035
Word decoding	.875	< .001	.802-.928	.035
Phonological processing	.849	< .001	.772-.907	.038
Writing efficiency	.734	< .001	.646-.810	.061
Reading comprehension	.743	< .001	.656-.818	.054
Vocabulary	.753	< .001	.667-.827	.054

Note. only participants who had accomplished the full screening protocol were entered into the analysis (N=121)

Table 6.

Pairwise comparisons of ROC curves for each test against the ROC curve for the entire protocol.

Test	*Z* statistics	Significance level	Difference between areas	Standard Error
Spelling	2.08	.037	.060	.029
Word-Chain	2.29	.022	.045	.020

Pseudo-homophone	2.14	.032	.071	.033
Writing Efficiency	3.43	.000	.186	.054
Reading Comprehension	3.91	.000	.177	.045
Vocabulary	3.82	.000	.166	.044