Identifying dyslexia in adults: an iterative method using the predictive value of item scores and self-report questions

Tamboer, P.; Vorst, H.C.M.; Oort, F.J.

DOI
10.1007/s11881-013-0085-9

Publication date
2014

Document Version
Final published version

Published in
Annals of Dyslexia

Ann. of Dyslexia (2014) 64:34–56
DOI 10.1007/s11881-013-0085-9

Identifying dyslexia in adults: an iterative method using the predictive value of item scores and self-report questions

Peter Tamboer · Harrie C. M. Vorst · Frans J. Oort

Received: 20 January 2013 / Accepted: 18 August 2013 / Published online: 21 December 2013

Abstract Methods for identifying dyslexia in adults vary widely between studies. Researchers have to decide how many tests to use, which tests are considered to be the most reliable, and how to determine cut-off scores. The aim of this study was to develop an objective and powerful method for diagnosing dyslexia. We took various methodological measures, most of which are new compared to previous methods. We used a large sample of Dutch first-year psychology students, we considered several options for exclusion and inclusion criteria, we collected as many cognitive tests as possible, we used six independent sources of biographical information for a criterion of dyslexia, we compared the predictive power of discriminant analyses and logistic regression analyses, we used both sum scores and item scores as predictor variables, we used self-report questions as predictor variables, and we retested the reliability of predictions with repeated prediction analyses using an adjusted criterion. We were able to identify 74 dyslexic and 369 non-dyslexic students. For 37 students, various predictions were too inconsistent for a final classification. The most reliable predictions were acquired with item scores and self-report questions. The main conclusion is that it is possible to identify dyslexia with a high reliability, although the exact nature of dyslexia is still unknown. We therefore believe that this study yielded valuable information for future methods of identifying dyslexia in Dutch as well as in other languages, and that this would be beneficial for comparing studies across countries.

Keywords Adult dyslexia · Classification · Criterion · Cross validation · Item scores Self-report

P. Tamboer (✉) · H. C. M. Vorst
Faculty of Social and Behavioural Sciences, Department of Psychology Methodology, University of Amsterdam, Weesperplein 4 Room 213, 1018 XA Amsterdam, The Netherlands
e-mail: petertamboer@zonnet.nl

F. J. Oort
Faculty of Social and Behavioural Sciences, Department of Educational Research, University of Amsterdam, Nieuwe Prinsengracht 130, 1018 VZ Amsterdam, The Netherlands

P. Tamboer · H. C. M. Vorst · F. J. Oort
Overtoom 247B, 1054 HW Amsterdam, The Netherlands

Introduction

Dyslexia is considered to be a developmental disorder that persists into adulthood, with estimates of prevalence varying between 3 and 18 %. Dyslexic children are usually relatively reliably identified with poor reading and spelling abilities despite adequate intelligence, motivation or schooling. However, the identification of dyslexic adults is usually problematic—for instance, because school records of reading and spelling abilities are not always available. The lack of objective methods for identifying adult dyslexia forces researchers to make arbitrary decisions about the process of selecting dyslexic and non-dyslexic participants. Any attempt to find an objective and standardized selection method for dyslexia is complicated by the development of an overwhelming quantity of theories and hypotheses of dyslexia in decades of research. In a recent study, Ramus and Ahissar (2012) discussed how the existence of so many theories complicates the interpretation of poor and normal performances of dyslexics in a broad range of tasks. They argued that any theory of dyslexia (e.g. phonological, visual, attentional) runs the risk of overgeneralizing poor performances—predicting poor performance in many more situations than observed—while normal performances are overlooked. For example, poor performances may be confounded by minor intellectual abilities or other factors. Normal performances on the other hand, may wrongly be interpreted as an indication of being not dyslexic, while the effects of dyslexia may just as well be too small to be detected in tasks that are too 'easy', especially in the case of people who are highly intelligent or people who underwent extra training during their school days or in college.

As a result of various interpretations of performances, researchers who need to select dyslexic and non-dyslexic adults are faced with various problematic issues, such as: which tests to use, which tests are considered to be the most reliable predictors of dyslexia, and how to determine cut-off scores. In a recent attempt to arrive at an objective method that deals with these issues, Tops et al. (2012) used a predictive analysis and found that the combination of only three tests (word reading, word spelling, and phonological awareness) sufficed for the identification of dyslexic and non-dyslexic students in higher education with more than 90 % accuracy. However, some issues in the process of identifying dyslexic adults remained unresolved in this study as well.

One issue is which exclusion and inclusion criteria should be applied to selected groups of dyslexics and non-dyslexics. For example, significant differences in general intelligence between groups are usually not accepted. The result is that in many studies, some participants—who are suspected of being dyslexic—are removed from further analyses because of low general intelligence. Aside from the disadvantage of the fact that dyslexic people with low intelligence are seldomly selected for studies of dyslexia, it is still unclear to what extent measurements of intelligence are affected by dyslexia in general.

A second issue is how to determine an objective criterion of dyslexia that can be used in a prediction analysis. The assumption in, for example, a discriminant analysis is that criterion groups exist beyond any doubt, so that a new sample of people can be classified based on the behaviour of these criterion groups in certain tasks. However, as long as we do not know exactly what causes dyslexia, we can never be absolutely sure of any individual being dyslexic or not. Even for very clear cases of dyslexia—individuals showing all known symptoms—alternative explanations such as low intelligence can never be ruled out completely. Realizing this, a criterion of dyslexia can never be determined with absolute certainty. An additional danger is that a criterion of dyslexia is based on the same kind of tasks as the tasks which are used in the predictive analysis. In the study of Tops et al., the dyslexic students were previously examined by a specialized remediation service and

retested if necessary. However, the authors did not report specific details about the tests which were used for a criterion of dyslexia, or how cut-off scores were determined.

A third issue is which tests should be administered. Tops et al. found that a combination of three different tests resulted in the most reliable prediction, which is consistent with the view that dyslexia is characterized by multiple deficits. However, some other cognitive impairments which have been described for dyslexics were not incorporated in the study of Tops et al. The number of tests needed for the best diagnosis of dyslexia remains unclear, but it seems safe to say that researchers who administer only a few tests for selection purposes run the risk of selecting false positives or false negatives because any single poor or normal performance does not necessarily mean that someone is dyslexic or not.

A fourth issue is how to determine which tests are the most important. This is a difficult issue for researchers who make use of an extended battery of tests. With the existence of so many theories of dyslexia, the risk exists that choices between tests are influenced by theoretical insights favoured by researchers. For example, one researcher might consider a poor performance on a phonological test to be the most important indication of dyslexia and a poor performance on a visual test less important, while another researcher might decide the other way around. This is not a reproach of researchers because as long as we do not know exactly what causes dyslexia, any decision can only be subjective. This subjectivity is strengthened by the fact that the performances of dyslexics on specific language tests vary widely between languages and across samples.

A fifth issue, related to the previous one, is how to determine cut-off scores for various tasks. This issue is complicated by the difficulty of interpreting poor and normal performances, as discussed by Ramus and Ahissar (2012). An additional complication is that it is unknown whether dyslexia is a distinct trait—just like handedness—or a disorder that varies in severity, representing the left-hand side of a normal distribution. Various theoretical views of dyslexia contribute to this issue. Some genetic and neurobiological findings support the view of one underlying deficit that develops differently, depending on the nature of schooling and training during childhood (e.g. Richlan, Kronbichler, & Wimmer, 2011). However, just as much support can be collected for the existence of subtypes of dyslexia (Bosse, Tainturier, & Valdois, 2007; Castles & Coltheart, 1993), which in turn can be distinct traits or traits that are normally distributed. Thus, to determine cut-off scores for cognitive measurements is a puzzling assignment for researchers leading to arbitrary choices.

A sixth issue is that the choice of which analyses to use is arbitrary. As an alternative to determining cut-off scores, discriminant analyses (DA) or logistic regression analyses (LR) can be used. As also explained by Tops et al., the general danger of prediction analyses is that the more predictors are used—which would be preferable for the identification of dyslexics—and the smaller the samples are, so a model fit will be based more and more on sample-specific variance. Considering this danger, the choice between DA and LR is not easy because the assumptions between these analyses differ considerably. For example, DA assumes normal distributions of explanatory data while LR does not. Pohar et al. (2004) evaluated both methods in various situations and concluded that DA should be preferred in some cases and LR in others.

In this study, we propose a method of diagnosing adult dyslexia which takes into account the above issues as much as possible. Thus, the aim was to identify dyslexic and non-dyslexic students with a high reliability and without having a preference for any theory of dyslexia beforehand. We took nine methodological measures, with most of them being new compared to previous methods. To address the first issue, we carefully considered several options for exclusion and inclusion criteria (e.g. intelligence, age, psychological health, fraudulent behaviour on tests). To address the second issue, we used six independent sources of biographical information for a criterion. We also varied the criterion by applying different decision rules for

the classification into the groups of dyslexics and non-dyslexics. In this way, we were able to compare predictions based on severe dyslexia with predictions based on moderate dyslexia. To address the third issue, we collected as many cognitive tests as possible, which together covered the entire spectrum of cognitive deficits in dyslexia. To address the fourth issue, we used two statistical techniques—DA and LR—which do not depend on subjective preferences for certain theories of dyslexia. To address the fifth issue, we not only used sum scores as predictor variables in the analyses, but also single item scores. As discussed by Ramus and Ahissar (2012), we assumed that dyslexic people do not necessarily make *more* mistakes than others, but that they make *specific types* of mistakes. These specific mistakes could be overlooked in prediction analyses using sum scores as predictors, but not, maybe, when item scores are used as predictors. To address the sixth issue, we did not choose between DA and LR beforehand. Instead, we compared the predictive power of the analyses with each other, by evaluating their assumptions for each separate prediction. Finally, to enhance the reliability of the predictions as much as possible, we took three additional measures. One additional measure was that we used a large sample of 495 Dutch first-year psychology students. With such a large sample, it was possible to use many predictors in the analyses, which were needed when using single items as predictors. A second additional measure was that we used self-report questions as predictor variables in the analyses. We assumed that this could enhance the reliability of predictions, because a self-report of very specific language difficulties might be less vulnerable to the influences of intelligence, schooling and compensation strategies. A third additional measure was that we repeated predictions on different criterion groups. We reasoned that when consistency between separate predictions is high for a participant, the reliability of the classification would be enhanced.

Method

Participants and sample characteristics

For this study, data of 1,110 first-year psychology students from the University of Amsterdam were used. This group is the summation of two separate groups of first-year students of 2009 (548 students) and 2010 (562 students). We excluded 615 students from any further analyses, more than half of the original group. At first sight this seems strange and may raise questions about the objectivity of procedures. Therefore, we will explain the exact procedures that resulted in this large exclusion of students (see "Procedure").

The remaining group of 495 students consisted of 125 males and 370 females. Mean age was 19.7 (1.5) years with a minimum age of 17.8 years, and a maximum age of 25.9 years. There were 414 right-handed students (84.5 %), 53 left-handed students (10.8 %), 23 ambidextrous students (4.7 %), and 5 students without handedness data.

Questionnaires

The *SES Questionnaire* (socio-economical status) is a questionnaire that is administered at the University of Amsterdam every year. In this study, we used information from this questionnaire about age, gender, nationality, language background, history of health problems, type of education, and school grades.

The *Handedness & Orientation Questionnaire*—designed for this study—aims to acquire all information about hand preference and orientation. We administered this questionnaire for two reasons. Years ago, many researchers assumed that dyslexia and left-handedness might be correlated. However, there is only a limited amount of empirical data available

regarding this correlation, with inconsistent results (Saviour et al., 2009; Van Strien, 1992). Nevertheless, at present, it is still common to report percentages of left-handedness in studies of dyslexia. The questionnaire consists of two parts. In the first part, questions are asked about hand preference in general (right-handed, left-handed, ambidextrous), hand preference in writing (right, left, both, right but as a child a preference for left), hand preferences of family members, and about 22 specific preferences (five response categories) such as: 'Which hand do you use for brushing your teeth?' The second part consists of ten statements about orientation between left and right in general (yes, no, a little) such as: 'As a child it was hard for me to determine the difference between left and right'.

The *Dyslexia Questionnaire*—designed for this study—aims to acquire general information that can be used to create a criterion of dyslexia. There are three parts: (1) a full self-report of dyslexia; (2) former test results from school or of an official institute of dyslexia; and (3) information about dyslexic family members (biological mother, father, sisters, or brothers).

The *Language Preference Questionnaire*—designed for this study—aims to acquire information about *general* language difficulties. The term 'dyslexia' is not mentioned in this questionnaire. There are six subscales, each consisting of ten statements with seven response categories. Here, we give an example of each subscale. *Reading*: 'Every week I read a book'; *Speaking*: 'I like to talk fast'; *Writing*: 'I keep a diary'; *Mental representations*: 'I remember faces easily'; *Memory*: 'I remember phone numbers easily'; *Foreign languages*: 'I don't like the fact that most study books are written in English'.

The *Communication Questionnaire* (Kramer & Vorst, 2007) aims to acquire information about *specific* language difficulties. The term 'dyslexia' is not mentioned in this questionnaire. The questionnaire is designed according to a 7×5×4 facet design. There are seven subscales representing different aspects of language, each consisting of 20 statements with seven response categories. Here, we give an example of each subscale. *Reading*: 'Sometimes I skip a letter, which results in reading a different word'; *Writing*: 'Sometimes I forget to write down a syllable'; *Speaking*: 'While speaking, I sometimes exchange similar words'; *Listening*: 'I hear a story exactly like someone tells it'; *Copying*: 'When I copy out a text, I sometimes exchange letters with similar sounds'; *Dictating*: 'I make mistakes in a dictation, because I don't hear the correct sounds'; *Reading aloud*: 'When reading aloud, I sometimes repeat a part of the text'. All statements can also be categorized into five subscales representing *sounds, letters, words, sentences and text*. A leading thought during the creation of statements was that four typical mistakes might distinguish dyslexics from others: *skipping (forgetting), adding, changing, and exchanging*.

Intelligence tests

Six cognitive tests were based on Guilford's Structure of Intellect Model: (1) *Vocabulary* (cognition of semantic units: knowing and understanding words and concepts); (2) *Verbal analogies* (cognition of meaningful verbal relations); (3) *Conclusions* (cognition of meaningful symbolic relations, or the ability to understand and structure difficult situations and the evaluation of semantic implications); (4) *Numeric progressions* (cognition of symbolic systems in progressions of numbers); (5) *Speed of calculation* (the ability to assess simple symbolic rules); and (6) *Hidden figures* (spatial intelligence). For general (non-verbal) intelligence we used *Advanced progressive matrices set 2* (Raven, Court, & Raven, 1979).

Short-term memory test

The *Short-term memory test*—designed for this study—aims to measure the capacity of short-term memory. We used the concept of digit span: the number of digits a person can

retain and recall. There are four subtests: numbers and letters, both forward and backward. And each subtest consists of 24 series: 6 of 4, 6 of 5, 6 of 6, and 6 of 7 items for the subtests numbers and letters forward, and 6 of 3, 6 of 4, 6 of 5, and 6 of 6 items for the subtests numbers and letters backward. The numbers and letters are presented one by one, for one second each on a computer screen. The participants have to retype these numbers and letters after the last one of a series has been presented. About half of all series consists of some typical difficulties for dyslexics, either phonological, visual, or both. For example, a typical phonological confusion is between the numbers seven and nine which resemble each other phonologically in Dutch (zeven/negen). Typical visual confusions are between the numbers six and nine and the letters [m] and [w]. The letters [p], [d], and [b] resemble each other phonologically as well as visually.

Specific language tests

Eleven (six auditory and five visual) language tests were designed for this study. We incorporated many typical difficulties for dyslexics in these tests. For the six auditory language tests, instructions and all test items were read out aloud in a well-trained female voice and can be heard through headphones and read on a computer screen at the same time. Item responses have to be typed within a limited time and can be started while listening. For all tests, high scores represent good performances and low scores represent poor performances.

Dutch dictation (auditory) aims to measure spelling abilities in the Dutch language (10 sentences, maximum score 10×4=40).

English dictation (auditory) aims to measure spelling abilities in the English language (10 sentences, maximum score 10×2=20). It can be assumed that Dutch students are familiar with the ordinary English words we used.

Missing letters (auditory) also aims to measure spelling abilities in the Dutch language (10 sentences, maximum score 10×2=20), but in a slightly different way. For each sentence, two words are repeated while these words are shown on the computer screen with a few letters left out of the word.

Pseudowords (auditory) aims to measure spelling abilities of pseudowords—non-words that sound like real words (30 words, maximum score 30). Participants have to decide whether the non-words they hear are spelled correctly on the computer screen. Usually, pseudowords are admitted the other way around through participants reading aloud the words themselves. The reason for changing this was that it would be practically impossible to get all students in private sessions for this way of testing.

Sound deletion (auditory) aims to measure phonological abilities (20 words, maximum score 20). Participants have to decide whether the difficult Dutch words they hear are pronounced correctly, and if not, which letter is missing or has been added (there is a choice between three words). For example the existing word ‘fietsenstalling’, which means bicycle shed, is read out as ‘fiestenstalling’. The possible answers are: ‘fietsentalling’, ‘fiestensalling’ and ‘fiestenstalling’.

Spoonerisms (auditory) (see also Hazan et al., 2009) also aims to measure phonological abilities (20 words, maximum score 20). A spoonerism is a word that consists of two existing smaller words and that still consists of two small existing words when the first letters of both small words are exchanged. For example, participants hear the word ‘kolen-schop’ which has to be altered to ‘scholen-kop’.

Incorrect spelling (visual) is the third test in our study that aims to measure spelling abilities in the Dutch language, again in a different way (40 words, maximum score 40). All

words are flashed on a computer screen for 50 ms. Participants have to decide whether the words are spelled correctly or not.

Dutch–English rhyme words (visual) aims to measure the ability to recognize similar-sounding nouns in Dutch and English (40 words, maximum score 40). Dutch–English word pairs are shown on a computer screen with the Dutch words on the right. Participants have to decide whether the words rhyme with each other or not. Typical confusion may arise in this test because the non-rhyming items have the same vowels, such as 'Deep-Reep'.

Letter order (visual) aims to measure the ability to read words as a whole (20 sentences, maximum score 20×2=40; time limit of 5 min). Theoretical hypotheses about reading words as a whole are described in *the dual route model of reading* (for an extended description see De Groot et al., 1994). The idea for this test comes from a well-known text: 'Aoccdrnig to rscheearch at Cmabrigde uinervtisy, it deosn't mttaer waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteres are at the rghit pclae. The rset can be a tatol mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae we do not raed ervey lteter by it slef but the wrod as a wlohe'. We created 20 sentences based on the same principle: the order of the letters of the words was changed, apart from the first and last letters. The words in the sentences are more difficult towards the end of the test. The sentences have to be typed in with all words correctly spelled. There are no words that consist of typical dyslexic spelling difficulties.

Counting letters (visual) aims to measure the effects of global reading (two sentences, maximum score 6+8=14). The idea for this test is based on a well-known language joke. Count the number of times that the letter [f] appears in the following sentence: 'Finished files are the result of years of scientific study combined with the experience of years'. Many people only see the [f] three times. It has been suggested that the [f] in [of] is overlooked because it sounds more like a [v]. Another suggestion is that [of] is overlooked completely as a result of global reading. We created two Dutch sentences based on the same principle. In the sentence 'Het deftige hondje van de man en de vrouw drinkt water uit de kraan' participants have to count the letter [d] (6). In the sentence 'Met de neus en de mond is het niet moeilijk en zelfs gemakkelijk een liedje te neuriën' participants have to count the letter [n] (8).

Mirror reading (visual) aims to measure the ability to read in mirror image (20 words, maximum score 20; 10 sentences, maximum score 10×4=40). On a computer screen, the words and sentences are presented in mirror image, so that they would appear normally written when held in front of a mirror. This means that the letters are also shown as a mirror image!

Reliability

The reliability of all tests and questionnaires was determined based on the sample of 495 students (see paragraph 2.3: "Procedure"). Table 1 shows the values of Cronbach's alpha. For *Letter order* and *Mirror reading* we did not calculate the reliability because the items were varying in difficulty and there was a time limit. For *Counting letters*, only one score was acquired. The majority of students performed on all tests and questionnaires. A relatively large group of 26 students did not have a score on *Spoonerisms* because they did not understand the instructions of the test.

Procedure

Collection of data

All data were collected at the University of Amsterdam during five sessions of 3 h each, in which tests and questionnaires of various studies were administered. These sessions took place

Table 1 Reliabilities of (sub-) tests and (scales of) questionnaires (Cronbach's alpha)

(Sub-) test/(Scales of) questionnaire	*N* items	*N* subjects	Cronbach's α
Language Preference Questionnaire	60	489	0.721
Handedness & Orientation Questionnaire	22	489	0.964
Communication Questionnaire (QC)	141	494	0.984
QC subscale reading	20	494	0.892
QC subscale writing	20	494	0.902
QC subscale speaking	20	494	0.905
QC subscale listening	20	494	0.910
QC subscale copying	20	494	0.942
QC subscale dictating	21	494	0.939
QC subscale reading aloud	20	494	0.942
Short-term memory numbers forwards	24	487	0.767
Short-term memory numbers backwards	24	482	0.804
Short-term memory numbers total	48	480	0.863
Short-term memory letters forwards	24	492	0.773
Short-term memory letters backwards	24	485	0.815
Short-term memory letters total	48	485	0.873
Dutch dictation	10	494	0.670
English dictation	10	493	0.425
Missing letters	10	490	0.627
Pseudowords	30	489	0.518
Sound deletion	20	489	0.847
Spoonerisms	20	469	0.906
Incorrect spelling test	40	493	0.584
Dutch–English rhyme words	40	493	0.788
Letter order	20	492	x
Counting letters	24	482	x
Mirror reading words	20	492	x
Mirror reading sentences	10	492	x

on midweek evenings with one or two weeks between each session. All students were obligated to participate because these sessions were part of the first-year study programme. All students were informed about the general nature of the tests and questionnaires in advance according to a standard protocol. However, regarding the tests and questionnaires related to dyslexia neither the students nor the surveillants knew about the true purpose. Afterwards, the students received a more detailed debriefing. Anonymity was guaranteed by the standard protocol of the University of Amsterdam. All students had up to 3 weeks after their debriefing to request that their test results were not used.

We made sure that our tests and questionnaires were spread out as much as possible, ensuring that they were not administered in one session but over all sessions and positioned between tests of other studies. We were also able to make sure that the order of tests was about the same in both years. However, we had no influence on the positioning of the standard intelligence tests. On our website (www.vorstmulder.nl), 39 test results can be found for the 2009 and 2010 groups. A few significant differences were found, but not more than could be expected when performing many separate *t* tests. However, most intelligence

tests were performed better by the first year which might be due to differences in the positioning of these tests (different sessions or different positions within one session).

Exclusion of participants

Of 1,110 participants 615 were excluded from any further analyses, leaving 495 students. We performed this process of exclusion in three successive steps.

In the first step, we excluded 361 students on three grounds: health, nationality, and missing data, leaving 749 students. A small group of students had a history of moderate to severe head trauma or severe general health problems (3 %). A large group of students was not 'completely Dutch' (25 %). This group consisted of foreign students, students who had not lived in The Netherlands for their entire youth, and students with one foreign parent (only when they indicated Dutch not as their primary mother language). Also excluded were students who did not participate in a large number of tests and questionnaires (10 %). There was quite some overlap between the last two groups, which makes sense because, for example, most foreign students did not participate in the language tests.

In the second step, we excluded 70 students aged between 26 and 54 years, leaving 679 students aged from 17 to 25 years. We excluded older students because we wanted our sample to be as homogeneous as possible, so that effects resulting from differences in long-term experience with language could be prevented.

In the third step, we excluded 184 students who we suspected were not serious in a few of the tests or questionnaires which were crucial for our study (27 % of 679). This seems a lot. However, we realized that the test sessions—which lasted three hours each—required much concentration and could have been exhausting for some students, leading to rushing through tests or questionnaires. After excluding a small group of very obvious cases of fraudulent behaviour, we applied an objective procedure to detect as much fraudulent behaviour as possible. First, we applied several statistic procedures that can detect answer patterns. Second, we excluded students whose registered test durations were so short that it would be theoretically impossible to even read the question in such a short time. This was, in many cases, probably due to test instructions that were not well understood. Finally, we decided that students with at least four fraudulent results in separate tests or questionnaires were excluded. In the remaining cases, we only removed the scores of single tests or questionnaires.

We did not exclude participants based on intelligence scores for two reasons. First, in a sample of students, it can be assumed that general intelligence is within the normal range. Second, it is not clear to what extent specific measurements of intelligence correlate with dyslexia.

Results

In the first paragraph, we describe how we determined the criterion groups of dyslexics (D-group) and non-dyslexics (ND-group). In the second paragraph, we describe the procedure of identifying dyslexic and non-dyslexic students. In the third paragraph, we describe various characteristics of the D-group and the ND-group by comparing these groups with the original criterion groups and by comparing test results between groups.

Criterion and classification

Ideally, when using prediction analyses, a perfect criterion is required. This criterion separates two small groups of people so that other subjects can be classified based on the behaviour of the

criterion groups in certain tests. To determine a perfect criterion of dyslexia is of course impossible when the exact nature of dyslexia is unknown. The only thing we can do is try to approximate a perfect criterion, which will always be based on arbitrary decisions.

In studies of dyslexia in the Netherlands, dyslexics are often selected based on the so called 'dyslexieverklaring', an official document that can be acquired by children after an extended testing procedure that is performed by a specialized educational psychologist. Although this document is widely accepted to be a reliable criterion of dyslexia in the Netherlands, it has three drawbacks. First, it cannot be ruled out that—where there is some doubt—some psychologists will decide to give a child a 'dyslexieverklaring'. Second, the reliability of this document suffers from various issues mentioned in the introduction. For instance, different psychologists use qualitatively different tests, a different number of tests and/or subjective standards in the choice of cut-off scores. Third, the use of the 'dyslexieverklaring' as a criterion of dyslexia for further research is not appropriate when the same kind of tests are investigated as were used for the criterion. In other words: the predictive power of a test can never objectively be investigated when a similar test was used in a criterion.

In this study, we took three approaches for improving the reliability and objectivity of a criterion. First, alongside the 'dyslexieverklaring', we used five other independent sources of biographical information as indications of dyslexia. Second, these other sources of information were not based on test results, so that we would not compare predictions and a criterion that were based on the same information. And third, we used different criterion groups, starting with strict criterion groups and then adjusting them based on information from various predictions. The advantage of this was that we could compare predictions based on the behaviour of severely impaired dyslexics and 'super controls' with predictions based on the behaviour of groups which included mildly impaired dyslexics and poor-performing non-dyslexics. The use of additional information from the predictions itself for the criterion is of course a statistical pitfall, especially because we reasoned that we should be careful in using the same information twice. However, we also reasoned that this would be justified as long as not too many false positives and false negatives were detected. An additional advantage is that the reliability of predictions would increase in the case of high consistency between predictions, just as when applying cross-validation procedures. For the first criterion groups, we only used biographical indications of dyslexia which were acquired from the *Dyslexia Questionnaire*. The indications and decision rules for classification are presented below. Students with inconsistent criteria were classified into a separate group.

Indications of dyslexia:

1. *Formal diagnosis of dyslexia by a qualified educational psychologist (yes, no).*
   In the Netherlands, the so called 'dyslexieverklaring' is an official written document that a dyslexic child can acquire after an extensive individual test session performed by a qualified educational psychologist. Despite some drawbacks, as mentioned above, we consider this document as a *strong indication* of dyslexia in this study.
2. *Other test results from school or dyslexia institute (dyslexic, non-dyslexic, maybe-dyslexic).*
   Until recently, the cost of acquiring an official document of dyslexia was high. Therefore, it was common in many schools to arrange a testing procedure within the school. These test results are less reliable than an official document, but still provide important information about the history of language difficulties of an individual student.

We considered a positive result ('dyslexic') and an indication of doubt ('maybe-dyslexic') as a *strong indication* of dyslexia.

3. *Extra language lessons in primary school (yes, no).*
   Some students were never officially or unofficially tested for dyslexia, but reported a history of extra language lessons during childhood. We believe that any sign of difficulties with language should be incorporated into a criterion. However, we considered extra lessons as a *weak indication* of dyslexia because language difficulties might just as well be the result of general learning difficulties. On the other hand, the importance of this indication should not be underestimated in our study, because it can be expected that most university students do not have general learning difficulties.
4. *Dyslexia in family (mother, father, brothers, sisters) (yes, no, not sure).*
   Heritability estimates of dyslexia in twin studies vary between 40 and 80 % (Hensler et al., 2010; Scerri & Schulte-Körne, 2010; Schumacher et al., 2007). A child with a dyslexic parent has a 40–60 % chance of developing dyslexia, a risk that increases if there are more dyslexic family members (Ziegler et al., 2005). These percentages clearly show the usefulness of information about family members. We considered having one or more dyslexic family member as a *strong indication* of dyslexia, and being not sure about having one or more dyslexic family member as a *weak indication*.
5. *General self-assessment of dyslexia (no, maybe, moderate, severe).*
   A general self-assessment of dyslexia consisted of only one question: 'Are you dyslexic?' We considered the answers 'severe' and 'moderate' as *strong indications* of dyslexia and 'maybe' as a *weak indication*. We assumed that this indication is relatively reliable, because most students in the Netherlands know what dyslexia is. It can therefore be expected that highly educated students in particular—who have just finished school—are capable of assessing for themselves to what extent they were different from other children. In other words: they probably know that something was wrong when they experienced difficulties performing tasks which were easy for most of the other children.
6. *Specific self-assessment of dyslexia: five statements (yes, no, a little).*
   There are five statements: (1) 'As a child, I experienced difficulties with reading'; (2) 'As a child, I experienced difficulties with spelling'; (3) 'I still experience difficulties with reading'; (4) 'I still experience difficulties with spelling'; and (5) 'I experience difficulties with writing'. The answers were scored as follows: yes (2), a little (1) and no (0). After adding up the five scores, the total score ranged from 0 to 10 with 10 representing severe difficulties with language. We marked scores of six or higher as a *strong indication* of dyslexia and a score of three, four or five as a *weak indication*.

Decision rules and classification in criterion groups:

1. *Dyslexic (D-group) (N=33).*
   We required the following: a formal diagnosis of dyslexia, together with at least three strong or two strong and two weak indications of dyslexia.
2. *Non-dyslexic (ND-group) (N=256).*
   We required the following: no formal diagnosis of dyslexia, no strong indications and not more than one weak indication of dyslexia.

   Of the 495 students, 206 did not satisfy the strict requirements to be classified into either of the two criterion groups. Thus, for these students indications of dyslexia were inconsistent. Most of these students had one strong or more than one weak indication of dyslexia. Three students were of special interest, because they had a formal diagnosis of

dyslexia, but not the number of indications needed to be classified into the D-group. As mentioned before, we accounted for the possibility that some students might have received a formal diagnosis on false grounds.

Table 2 shows the number of times that various indications are represented in the two criterion groups. In the ND-group, only 32 weak indications were reported (all for different students), mostly regarding family members who might be dyslexic. Thus, 77.5 % of this group did not have any indication of dyslexia. In the D-group all students reported a formal diagnosis, which was required. Strong indications derived from former test results, general self-assessment, and specific self-assessment, and are common. And about half of these students reported dyslexic family members.

Predictions

The aim of this study was to identify dyslexic and non-dyslexic students with a high reliability. This can be accomplished by classifying as much as possible students with high consistency between the criterion and separate predictions. We therefore carried out predictions based on different predictor variables (sum scores, factor scores and single item scores of tests and questionnaires) and repeated the predictions using criterion groups which were acquired after intermediate adjustments. After each round of predictions, we classified all students based on decisions which are described below.

Sum scores and factor scores

For sum scores of questions, we determined the total scores of seven subscales of the *Communication Questionnaire* (CQ): reading, writing, speaking, listening, copying, dictating, and reading aloud. Attempts to arrive at meaningful factor scores failed as a result of the large quantity of items (20 for each subscale).

For sum scores of tests, we determined the total scores of 13 tests: specific language tests and *short-term memory* (STM) *letters and numbers*. After an exploratory factor analysis, these 13 sum scores loaded on four components which together explained 54 % of all variance. The factor scores of these components were used for a separate prediction. Appropriate names for these factors are 'Spelling' (*Dutch dictation*, *English dictation*, *Missing letters*, and *Incorrect spelling*), 'Phonology' (*Pseudowords* and *Sound deletion*), 'STM' (*STM numbers and letters*) and 'Order' (*Letter order*, and *Mirror reading*). The tests *Spoonerisms*, *Dutch–English rhyme*

Table 2 Criterion groups: specific indications of dyslexia

Indication of dyslexia	Dyslexic (*N*=33)		Non-dyslexic (*N*=256)	
Former diagnosis (required for D-group)	33	100 %	0	0 %
Former test result (strong indication)	31	94 %	0	0 %
Extra lessons (weak indication)	21	64 %	3	1 %
Dyslexia in family (strong indication)	16	48 %	0	0 %
Dyslexia in family (weak indication)	11	33 %	25	10 %
Self-assessment general (strong indication)	32	97 %	0	0 %
Self-assessment general (weak indication)	1	3 %	2	1 %
Self-assessment specific (strong indication)	27	82 %	0	0 %
Self-assessment specific (weak indication)	5	15 %	2	1 %

words, and *Counting letters* were involved in various factors. Table 3 shows the factor loadings of each test in a rotated component matrix.

Item scores

The advantage of using single items as predictor variables instead of sum scores is that they account for the possibility that the mistakes dyslexic people make are qualitatively different from 'normal' mistakes. We could, for example, imagine a highly intelligent dyslexic person who performs relatively well on a language test, but who makes mistakes on most of the typical dyslexic items. This is an example of 'overlooking normal performances', as described by Ramus and Ahissar (2012). The reverse situation would be a non-dyslexic person with low intelligence who performs relatively poorly on a language test, but not necessarily on the typical dyslexic items. This person might falsely be diagnosed as dyslexic using traditional diagnosing methods.

All available questions were used as predictors in two separate analyses, 141 questions from the CQ and 60 questions from the *Language Preference Questionnaire* (LPQ). Furthermore, 242 items of 10 tests were selected as predictors in a third analysis. We did not select items of *Mirror reading* because the performance on this test depended merely on speed of performance, nor of the STM tests because the number of items of these tests would enlarge the total number of items too much. For *English dictation* and *Missing letters*, both sentences as separate words were used as predictors (every item was a sentence that consisted of two sub-items which were words).

The danger of many predictor variables is that a model fit will for a large part be based on sample-specific variance. In this study, the number of selected items exceeded the number of cases in the smallest group, the D-group (both the original and the adjusted criterion D-group). Therefore, we selected only a relatively small number of items for a third round of predictions. The procedures we used to acquire this selection are described in a follow-up study (in preparation). In this study, we used two subsets of items for two predictions: a set

Table 3 Rotated component matrix of 13 tests

Test	Component			
	Spelling	Phonology	STM	Order
Dutch dictation	0.73	0.18	0.05	0.20
English dictation	0.62	0.08	0.15	0.08
Missing letters	0.65	0.18	0.08	0.23
Pseudowords	0.16	0.69	0.13	0.23
Sound deletion	0.01	0.81	0.10	−0.12
Spoonerisms	0.26	0.11	0.10	0.35
Incorrect spelling	0.73	−0.11	0.17	0.14
Dutch–English rhyme words	0.25	0.10	0.42	0.31
Letter order	0.12	0.05	0.10	0.66
Counting letters	0.51	−0.11	0.35	−0.27
Mirror reading	0.07	−0.06	0.16	0.72
Short-term memory numbers	0.14	0.14	0.85	0.16
Short-term memory letters	0.16	0.12	0.86	0.16

Extraction method: principal component analysis

Rotation Method: Varimax with Kaiser Normalization

that contained 88 test items and a set that contained 80 questions of the CQ, LPQ, and the *Orientation Questionnaire* (OQ).

Characteristics of discriminant analysis and logistic regression analysis

We used the *stepwise method* of DA and LR to predict group membership (dyslexic or non-dyslexic), because we wanted to acquire a reduced set of predictors while we did not want to assign some predictors higher priority than others. We entered the criterion groups as criterion in the analysis. For the DA, we set prior probabilities to *all groups equal*. A cross-validation procedure was chosen by selecting the option *leave-one-out classification* in order to prevent a single case being classified partly based on its own behaviour, and to correct for the effects of outliers. All analyses were carried out in SPSS.

Choices between predictions in this study

For all predictions, we carried out both DA and LR. Pohar et al. (2004) evaluated DA and LR in various situations. We considered their recommendations to be a reliable guideline for evaluating various predictions. According to these researchers DA should be preferred over LR in the case of small sample sizes, when violations of normality of predictor variables are not too bad (skewness within the interval [−0.2, 0.2]) and in the case of categorical explanatory variables with four or more answer categories. Based on these recommendations we made a choice between DA and LR for each separate prediction.

1. For a prediction based on the sum scores from the CQ, DA should be preferred because the scores (except one) were distributed with skewness within the interval [−0.2, 0.2].
2. For a prediction based on the sum scores from the tests, LR should be preferred because most predictors were distributed with skewness outside the interval [−0.2, 0.2] (but within the interval [−0.4, 0.4]).
3. For a prediction based on the single questions from the CQ, DA should be preferred because all items had seven answer categories with most of them showing a relatively limited skewness of distribution.
4. For a prediction based on the item scores from the tests, LR should be preferred because most test items were dichotomous.
5. For a prediction based on the single questions from the LPQ, DA should be preferred because all items had five answer categories, and most items were normally distributed.
6. For a prediction based on the factor scores from the tests, LR should be preferred because the factor scores were distributed with skewness below −0.5.
7. For a prediction based on a selection of questions from the CQ and LPQ, DA should be preferred because all items had five or seven answer categories with most of them showing a relatively limited skewness of distribution.
8. For a prediction based on a selection of item scores from the tests, LR should be preferred because most test items were dichotomous.

Decision rules

Two of the steps that we took to acquire reliable predictions were that we made sure that various predictions were independent and that we repeated the predictions on adjusted criterion groups. In this way, we expected to acquire test–retest reliability, with the additional benefit that predictions would not only be based on the behaviour of extreme groups of

dyslexics and non-dyslexics, but also on the behaviour of dyslexics who only show mild impairments. After various predictions, the criterion groups were adjusted following certain decision rules. All students were reassigned six times and this mostly affected the students who were not classified into the first criterion groups. Each step of reassignment is described below. Table 4 shows the results of all these steps.

Reassignment 1: We adjusted the original criterion groups based on six predictions (sum scores CQ and tests; factor scores tests; item scores CQ, LPQ and tests), so that the consistency between the criterion and the predictions was high. First, we added up the six predictions which resulted in a score that ranged from zero (low chance of dyslexia) to six (high chance of dyslexia). Next, we reassigned all students. The requirements for the new criterion of dyslexia were a prediction score of four or higher and not a previous criterion of non-dyslexia. The requirements for the new criterion of non-dyslexia were a prediction score of zero and not a previous criterion of dyslexia. The main result of this reassignment was that the D-group was enlarged.

Reassignment 2: We repeated all predictions with the new criterion groups, and repeated exactly the same reassigning procedure, so that we again acquired new criterion groups. The main result was that the group of dyslexics was enlarged even more. This might be explained by the fact that the predictions were now also based on the behaviour of moderately impaired dyslexics.

Reassignment 3: Another possibility is to add up all predictions of the two previous reassignment procedures. This resulted in scores that ranged from zero to 12. We determined a cut-off score of eight or higher for the new D-group, and zero for the new ND-group. Thus, at this point, we had evaluated the original criterion groups based on two separate reassignment procedures and on a prediction score of these two procedures together. Only with complete consistency between these three different classifications and following the same requirements regarding consistency with the original criterion, students were classified as dyslexic or as non-dyslexic. The group of dyslexics now became smaller again.

Reassignment 4: Next, we carried out predictions based on two selections of items (questions and test items). We added up these predictions. A score of two represented consistency regarding dyslexia and a score of zero represented consistency regarding non-dyslexia. The requirements for the new criterion of dyslexia were a prediction score of two and not a previous criterion of non-dyslexia. The requirements for the new criterion of non-dyslexia were a prediction score of zero and not a

Table 4 Reassignment (RA) procedures applied on criterion groups

	Criterion	RA1	RA 2	RA 3	RA 4	RA 5	Final RA
Dyslexic	33	49	73	46	62	71	74
Non-dyslexic	256	255	237	227	322	347	369
Remaining students	206	191	185	222	111	77	52

Table 5 Intermediate predictions of the original criterion group of dyslexics (*N*=33), leading to a final identification

Predicted as	Predictions first round						Predictions second round						Predictions third round		Predictions fourth round		Final identification
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
Dyslexic	25	16	29	23	29	16	26	20	29	19	26	18	31	25	30	26	33
Non-dyslexic	6	9	3	5	4	7	5	5	3	9	7	5	1	1	2	0	0
Missed predictions	2	8	1	5	0	10	2	8	1	5	0	10	1	7	1	7	0

previous criterion of dyslexia. The main result of this reassignment was that both the D-group and the ND-group was enlarged.

Reassignment 5: After the last procedure, 111 students were still not yet classified. For these students, we compared the last predictions with all previous ones again, and classified students in the case of high consistency. This was the case for 34 students. They had inconsistent predictions according to reassignment 4 but consistent predictions according to reassignment 3.

Final reassignment: We repeated the exact procedures of reassignment 4 and 5 on the new groups. Thus, we repeated the two predictions based on single items (as in reassignment 4), and again we added up these predictions. A score of two represented consistency regarding dyslexia and a score of zero represented consistency regarding non-dyslexia. Also students with a score of one but with consistent predictions before (as in reassignment 5), were classified as dyslexic or non-dyslexic. After this procedure, only 52 students remained without a reliable classification.

Predictions

Below, all predictions are summarized. Sum scores, factor scores and items which were inserted in the regression equation are reported as well. The items are not specified (for this, see our website www.vorstmulder.nl). Tables 5 and 6 show how the original criterion groups were predicted. For most intermediate predictions a few students could not be predicted because of missing variables. The number of students with missing predictions varied across

Table 6 Intermediate predictions of the original criterion group of non-dyslexics (*N*=256), leading to a final identification

Predicted as	Predictions first round						Predictions second round						Predictions third round		Predictions fourth round		Final identification
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
Dyslexic	25	2	3	10	12	2	34	10	10	16	24	4	7	9	4	1	1
Non-dyslexic	231	234	253	236	244	232	222	226	246	230	232	230	249	235	251	243	255
Missed predictions	0	20	0	10	0	22	0	20	0	10	0	22	0	12	1	12	0

predictions because, for each prediction, the predictor variables that were inserted in the equation differed.

Predictions first round:	
1. Sum scores CQ (DA)	Reading, Writing, Listening, Dictation
2. Sum scores tests (LR)	Dutch dictation, English dictation, Missing letters, Spoonerisms, STM numbers
3. Questions CQ (DA)	26 items
4. Item scores tests (LR)	12 items
5. Questions LPQ (DA)	13 items
6. Factor scores tests (LR)	Spelling, Memory, Order
Predictions second round (the same predictions of the first round repeated):	
7. Sum scores CQ (DA)	Reading, Writing, Listening, Dictation, Reading aloud
8. Sum scores tests (LR)	Dutch dictation, English dictation, Missing letters, Spoonerisms, STM numbers
9. Questions CQ (DA)	20 items
10. Item scores tests (LR)	8 items
11. Questions LPQ (DA)	12 items
12. Factor scores tests (LR)	Spelling, Memory, Order, Phonology
Predictions third round:	
13. Selection questions (DA)	15 items
14. Selection test items (LR)	13 items
Predictions fourth round (the same predictions of the third round repeated):	
15. Selection questions (DA)	15 items
16. Selection test items (LR)	24 items

Conclusion

From Table 4, it becomes clear that all except 52 students could be classified with high consistency. The original D-group was enlarged from 33 to 74, and the original ND-group was enlarged from 256 to 369. For 15 of the 52 not classified students, too many missed predictions made a reliable classification impossible. For the other 37 students, predictions were too inconsistent for a reliable classification.

Table 5 shows that all dyslexic students according to the original criterion were identified as dyslexic. Table 6 shows that all non-dyslexic students according to the original criterion were identified as non-dyslexic except for one. It should be noted that the false positives and false negatives according to the separate predictions were in most cases not the same students. Therefore, we can conclude that the false positives and false negatives of the separate predictions are mainly shortcomings of these predictions themselves, while the summation of repeated predictions is consistent with an independent criterion.

We further should note that most of the 37 students, who could not be classified, could also not be classified in a majority of the six reassignment procedures. A majority of 21 of these students could not be classified in all reassignment procedures; 10 students were classified as dyslexic once; one student was classified as non-dyslexic once; four students were classified as dyslexic three times; and one student was classified as dyslexic five times. Thus, only five students could be classified more than one time.

Another conclusion from Tables 5 and 6 is that the predictions based on a selection of items are the most reliable, because these predictions resulted in the smallest number of false positives and negatives compared to the original criterion groups.

Differences between groups

In paragraph 3.1, the number of times that various indications of dyslexia are represented in the two original criterion groups was presented in Table 2. Table 7 is the same as Table 2, but now with the final groups of dyslexics and non-dyslexics. A remarkable result is that two students with a formal diagnosis were classified as non-dyslexic, whereas it might be expected that predictions would be inconsistent for these students. Apparently, the former diagnosis was provided to the students on false grounds. The remaining percentages all make sense. We should note that the percentage of dyslexic students who have a strong indication of dyslexic family members is about the same for the criterion D-group (48 %) as for the final classification D-group (43 %). This indication can be considered to be the most objective one because it does not depend on a self-assessment or on an assessment by others. Therefore, this supports the reliability of the classification of dyslexics that do not have a former diagnosis.

Group differences of sample characteristics (gender, age and handedness) and sum scores of tests and questionnaires were calculated between the criterion groups and between the final groups. Furthermore, we wanted to know whether differences exist between the dyslexics according to the original criterion and the dyslexics that were added after the predictions, and likewise, between the non-dyslexics according to the original criterion and the non-dyslexics that were added after the predictions. This resulted in three large tables which can be found on our website (www.vorstmulder.nl). Here, we only present a summary of the results. The significance of differences between the various groups was determined with post hoc comparisons of an ANOVA analysis (Tukey, 0.05 level).

First, we calculated the differences between the three original criterion groups of dyslexics (33), non-dyslexics (256) and students without a criterion (206). Sample characteristics did not differ between the groups. The ND-group performed better than the D-group (most $p<0.005$) on all subscales of the CQ, on all STM tests, on all specific language tests (except *Sound deletion*), on three school grades (Dutch, English and other languages), and on the intelligence tests *Vocabulary*, *Verbal analogies*, *Speed of calculation*, and *Hidden figures*. No significant differences were found for *Raven progressive matrices*, *Conclusions*,

Table 7 Final classification results: specific indications of dyslexia

Indication of dyslexia	Dyslexic (*N*=74)		Non-dyslexic (*N*=369)	
Former diagnosis	33	45 %	2	1 %
Former test result (strong indication)	36	49 %	4	1 %
Extra lessons (weak indication)	36	49 %	16	4 %
Dyslexia in family (strong indication)	32	43 %	43	12 %
Dyslexia in family (weak indication)	13	18 %	31	8 %
Self-assessment general (strong indication)	33	45 %	0	0 %
Self-assessment general (weak indication)	19	26 %	15	4 %
Self-assessment specific (strong indication)	49	66 %	4	1 %
Self-assessment specific (weak indication)	13	18 %	32	9 %

Numeric progressions, and two other school grades (mathematics and remaining courses). All performances of the students without a criterion were higher than those of the D-group and lower than those of the ND-group, with some of these differences significant and some not significant.

Second, we calculated the differences between the two final groups of dyslexics (74), non-dyslexics (369) and non-identified students (37). Again, sample characteristics did not differ between the groups. The ND-group performed better than the D-group (all $p<0.0005$) on all subscales of the CQ, on all STM tests, on all specific language tests, on three school grades (Dutch, English and other languages) (all $p<0.005$), and on the intelligence tests *Vocabulary*, *Verbal analogies*, *Speed of calculation*, *Hidden figures* (all $p<0.0005$), *Numeric progressions* ($p=0.021$), and *Raven progressive matrices* ($p=0.022$). No significant differences were found for *Conclusions* and two other school grades (mathematics and remaining courses). The performances of the students that could not be identified (NI-group) were not, as might be expected—like those of the students without a criterion—somewhere in between the performances of the D-group and the ND-group. Instead, the performances of the NI-group on the specific language tests were sometimes close to the performances of the D-group and sometimes close to those of the ND-group. Furthermore, the performances on the subscales of the CQ were almost exactly the same as those of the D-group, being lower than those of the ND-group ($p<0.0005$).

Third, we calculated the differences between the original criterion group of dyslexics (33) and the later-identified dyslexics (41), and between the original criterion group of non-dyslexics (255, minus the false negative!), and the later-identified non-dyslexics (114). Sample characteristics did not differ between the groups. The two groups of dyslexics only differed on *Sound deletion* with the later-identified dyslexics having lower scores than the original dyslexics ($p=0.036$). But statistically, the finding of only one significant difference out of 39 comparisons is not relevant. The two groups of non-dyslexics only differed on *Incorrect spelling* with the later-identified non-dyslexics having lower scores than the original non-dyslexics ($p=0.042$). This is again statistically not relevant. However, the later-identified non-dyslexics also performed worse than the original non-dyslexics on all scales of the CQ (all $p<0.05$). Cross-comparisons between the two groups of dyslexics and the two groups of non-dyslexics for these last results showed that the performances of the later-identified non-dyslexics were still significantly better than of both groups of dyslexics (all $p<0.05$).

Discussion

With a new method of diagnosing dyslexia in adults, 89.5 % of 495 students could be identified as dyslexic (74) or non-dyslexic (369) with high reliability. The main characteristic of this method is that it is iterative: a criterion of biographical information was adjusted after repeated predictions. The most remarkable finding was that the most reliable predictions were acquired using test items (24) and self-report questions (15) as predictor variables instead of sum scores. We will discuss the reliability of the method, and the pros and cons of the method together with recommendations for future research.

Reliability of results

The main support for the reliability of the classification method is that there was a high consistency between separate predictions, which can be interpreted as an extended cross-

validation procedure. However, one might argue that with this method only sample-specific features were successfully identified, while it is uncertain whether these features are symptomatic of dyslexia. A first objection to this is that it is not known what the most characteristic features of dyslexia are in general. Furthermore, we believe that in this study the reliability of the classification method is also supported by the fact that the classification results were highly consistent with an independent criterion and by the fact that the predictions based on test results and the predictions based on a self-report of dyslexia were highly consistent with each other, especially when single test items and questions were used as predictor variables.

That the predictions based on items were the most reliable ones in our study was what we expected, especially after reading the discussion of Ramus and Ahissar (2012) regarding the interpretation of poor and normal performances. Sum scores are vulnerable to the overgeneralizing of certain mistakes, while it is unclear which mistakes are representative for the difficulties that accompany dyslexia. For instance, while phonological impairments are generally accepted, their exact nature remains unclear, with interpretations varying from deficits in the phonological lexicon or deficits in processes of retrieval from the lexicon (e.g. Blomert, Mitterer, & Paffen, 2004; Blomert & Willems, 2010). Thus, a collection of items can successfully distinguish dyslexics and non-dyslexics by identifying specific difficulties that accompany dyslexia. A requirement, of course, is that in the analysis items must be implied that together represent as many typical difficulties of dyslexia as possible. This was indeed the case for the two collections of predictor items. From the consistency between the predictions based on test items and the predictions based on self-report questions we conclude that, apparently, students are capable of identifying the exact nature of difficulties themselves, which is most remarkable when we realize that the questions were part of a questionnaire in which no link to dyslexia was made. We therefore conclude that the predictive power of self-report questions may have been underestimated so far. An extended description of the items is provided on our website (www.vorstmulder.nl). In a follow-up study (Tamboer et al., submitted), we used these collections of items for further analysis of symptoms and severity of dyslexia.

The fact that all 289 students with a strong criterion of dyslexia or non-dyslexia were classified correctly, except for one mismatch—a student without a criterion of dyslexia who was predicted as dyslexic—strongly supports the reliability of the method because the original criterion and the classification results were independent from each other. Two remarks should be made here. First, one might argue that there is no complete independency because the official document of dyslexia might have been partly based on the same type of tests as used in our prediction analyses. However, a requirement for a criterion of dyslexia was high consistency between biographical indications of dyslexia. In fact, we excluded three students with an official document from the criterion dyslexic group because other indications pointed to no dyslexia. Second, questions may arise about the independency between the biographical self-report questions of the criterion and the questions of the *Communication Questionnaire*. We emphasize here that these questions are both quantitatively and qualitatively different from each other. The biographical questions are merely questions that request information about a history of language-related difficulties, while the questions of the *Communication Questionnaire* request information about very specific difficulties in the use of language which are not necessarily related to dyslexia, and without mentioning that the questions were meant to measure dyslexics' difficulties.

More support for the reliability of the method can be derived from a comparison of characteristics and test results between the original criterion groups and the dyslexics and non-dyslexics that were identified with predictions only. For instance, the most objective indication of dyslexia in the criterion is family members being dyslexic. We found that 48 % of the criterion dyslexics and 40 % of the added dyslexics had a dyslexic parent, brother or sister, which is consistent with heritability estimates of dyslexia in twin studies varying between 40 and 80 % (Hensler et al., 2010; Scerri & Schulte-Körne, 2010; Schumacher et al., 2007). Furthermore, there were no differences between the two groups of dyslexics in performances of tests or subscales of questions, except for one, which was merely the result of making many post hoc comparisons. However, the later-identified non-dyslexics performed worse than the criterion non-dyslexics on all subscales of the *Communication Questionnaire* (but still better than the groups of dyslexics). An explanation might be that the later-identified non-dyslexics have overcome language difficulties which are not necessarily related to dyslexia. The fact the predictions based on questions for this group were consistent with other predictions underlines that using sum scores as predictors of dyslexia might be misleading.

Finally, a few remarks should be made about the differences between groups in measurements of intelligence. Generally, it is assumed that dyslexics and non-dyslexics do not differ in general intelligence. We found that the non-dyslexics in this study outperformed the dyslexics in some specific intelligence tests. For some tests, such as *Vocabulary* and *Verbal analogies*, this result makes sense because these tests partly depend on language skills. However, it was unexpected that the non-dyslexics also performed better than the dyslexics on the *Raven progressive matrices*, while no difference was found for *Conclusions*, both tests which are assumed to measure general intelligence. On the other hand, school grades were as expected, with lower school grades of language courses for dyslexics, but no differences in school grades of mathematical and other courses. An explanation for the difference in the *Raven progressive matrices* might be that performances on this test are actually influenced by dyslexia to some extent, but with this influence being so small that it usually remains undetected in small samples. This is an example of what Ramus and Ahissar (2012) mean with the underestimation of normal performances. We investigated this issue in a meta-analysis (in preparation).

Pros and cons of the method and recommendations

A shortcoming of the classification method in this study was that of 495 students 10.5 % could not be identified. For 15 students, this was due to too many missing variables. Before our analyses, we excluded many students because of fraudulent behaviour. However, in the case of students having only a few suspicious scores, only the suspicious scores were deleted. Apparently, this led to insufficient information for a reliable prediction for 15 students. Thus, of the remaining 480 students, 8 % could not be identified due to inconsistent predictions. One explanation is that the method in this study fell short as a result of using many predictor variables for some separate predictions. This might have resulted in the overestimation of dyslexics or non-dyslexics in separate predictions, leading to inconsistency between predictions. Another explanation is that there will always be a few dyslexics who manage to overcome some difficulties as a result of highly developed compensation strategies and a few non-dyslexics who just by chance perform poorly on some specific items or tests. A third explanation is that some students overestimate or underestimate their difficulties in the self-report questions,

which also leads to inconsistency between predictions. From a theoretical point of view, a fourth explanation might be derived from the unresolved issue as to whether dyslexia is a distinct trait or a trait which is normally distributed. Maybe there is a small group of people who only show a few symptoms of dyslexia, which might be caused by something else than dyslexia.

We took nine methodological measures with the aim to improve the reliability of diagnosing dyslexia in adults, with most of them being new compared to previous methods. Although most of them showed clear benefits, we found reason to believe that some measures can be improved in future studies, especially in studies with other samples than in our study. First, the advantage of applying strict exclusion and inclusion criteria for students to participate in further analyses was that it became clear that there might always be a group of people who cannot be identified for reasons other than differences in age, health, or intelligence. However, for samples of a general population instead of students, this might be a more complicated issue. Second, the criterion in this study was reliable as a starting point of further analyses. Third, the advantage of using many tests and questionnaires appeared to be that most of the symptoms were included in the most reliable predictions. However, one consideration would be to test for even more symptoms. For instance, attentional symptoms of dyslexia were mainly included in the self-report questions but not in the tests (while it cannot be excluded that some items of the tests depended partly on attention abilities). Fourth, the use of two statistical techniques (discriminant analysis and logistic regression analysis) which do not depend on subjective preferences for certain theories of dyslexia was successful by applying the *stepwise method*. In this way, we acquired a reduced set of predictors without assigning some predictors higher priority than others. Fifth, we avoided the use of cut-off scores by using items as predictor variables. These predictions proved to be more reliable and consistent than predictions based on sum scores. Sixth, instead of choosing between the analyses beforehand, a comparison between both analyses for each separate prediction resulted in more consistency between predictions than would have been the case using only one analysis. Seventh, using a large sample made it possible to use many predictors in the analyses, which is needed when using items as predictors. Eighth, using self-report questions as predictor variables supported the reliability of the identifications, because the resulting predictions were highly consistent with predictions based on test items. Ninth, the results showed that the consistency and reliability of predictions increased with the use of repeated predictions. An additional advantage of this study was that no productive tasks were used, which are generally time-consuming. For future methods of identifying dyslexia, this is an advantage in a practical sense.

Conclusion

This study showed that the reliability of diagnosing dyslexia in adults can be improved using an independent criterion in combination with using test items and self-report questions as predictor variables. The main characteristic of the analyses in this study was that the reliability could be improved with repeated predictions. We believe that this study provides new tools for diagnosing dyslexia for future studies, including in other languages. One clear disadvantage of the method in this study is that it requires a large sample and a long testing time. We have provided recommendations for future studies of which the most important one is that samples other than students' should be investigated. Details about this study can be found on our website (www.vorstmulder.nl).

Acknowledgments The authors thank Ineke van Osch for her help with the development of tests, Paul Brouwer and Nihayra Leona for their work and patience in programming many of our newly developed tests, and Jan Hoogeboom for his tremendous efforts in the processing of data.

References

Blomert, L., Mitterer, H., & Paffen, C. (2004). In search of the auditory, phonetic, and/or phonological problems in dyslexia: Context effects in speech perception. *Journal of Speech Language and Hearing Research, 47*, 1030–1047.

Blomert, L., & Willems, G. (2010). Is there a causal link from a phonological awareness deficit to reading failure in children at familial risk for dyslexia? *Dyslexia, 16*, 300–317.

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

Castles, A., & Coltheart, M. (1993). Varieties of developmental dyslexia. *Cognition, 47*, 149–180.

De Groot, A. M. B., Dannenburg, L., & Van Hell, J. G. (1994). Forward and backward word translation by bilinguals. *Journal of Memory and Language, 33*, 600–629.

Hatcher, J., Snowling, M. J., & Griffiths, Y. M. (2002). Cognitive assessment of dyslexic students in higher education. *British Journal of Educational Psychology, 72*, 119–133.

Hazan, V., Messaoud-Galusi, S., & Rosen, S. (2009). Speech perception abilities of adults with dyslexia: Is there any evidence for a true deficit? *Journal of Speech Language and Hearing Research, 52*, 1510–1529.

Hensler, B. S., Schatschneider, C., Taylor, J., & Wagner, R. K. (2010). Behavioral genetic approach to the study of dyslexia. *Journal of Developmental & Behavioral Pediatrics, 31*, 525–532.

Kramer, D. & Vorst, H.C.M. (2007). Communicatievragenlijst voor de opsporing van dyslectische leerlingen. Amsterdam. Thesis (Dutch). University of Amsterdam, Department of Psychology.

Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodološki, 1*, 143–161.

Ramus, F., & Ahissar, M. (2012). Developmental dyslexia: the difficulties of interpreting poor performance, and the importance of normal performance. *Cognitive Neuropsychology, 29*(1–2), 104–122.

Raven, J.C., Court, J.H., & Raven, J. (1979). A manual for Raven's Progressive Matrices and Vocabulary Tests. London: H.K. Lewis; San Antonio, Texas: The Psychological Corporation.

Richlan, F., Kronbichler, M., & Wimmer, H. (2011). Meta-analyzing brain dysfunctions in dyslexic children and adults. *NeuroImage, 56*, 1735–1742.

Saviour, P., Padakannaya, P., Nishanimutt, S., & Ramachandra, N. B. (2009). Familial patterns and biological markers of dyslexia. *International Journal of Human Genetics, 9*, 21–29.

Scerri, T. S., & Schulte-Körne, G. (2010). Genetics of developmental dyslexia. *European Child and Adolescent Psychiatry, 19*, 179–197.

Schumacher, J., Hoffmann, P., Schmäl, c., Schulte-Körne, G., & Nöthen, M. (2007). Genetics of dyslexia: the evolving landscape. *Journal of Medical Genetics, 44*, 289–297.

Tops, W., Callens, M., Lammertyn, J., Van Hees, V., & Brysbaert, M. (2012). Identifying students with dyslexia in higher education. *Annals of Dyslexia, 62*, 186–203.

Van Strien, J. W. (1992). Classificatie van links- en rechtshandige proefpersonen. *Nederlands Tijdschrift voor de Psychologie, 47*, 88–92.

Ziegler, J. C., Pech-Georgel, C., George, F., Alario, F.-X., & Lorenzi, C. (2005). Deficits in speech perception predict language learning impairment. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 14110–14115.