
Reading and Writing: An Interdisciplinary Journal **17:** 517–536, 2004.

Assessing reading skills by means of paper-and-pencil lexical decision: Issues of reliability, repetition, and word-pseudoword ratio

MARTINE A.R. GIJSEL, WIM H.J. VAN BON and ANNA M.T. BOSMAN
University of Nijmegen, The Netherlands

Abstract. This study focused on the feasibility of a group-administered paper-and-pencil lexical-decision test as a plausible alternative or supplementary tool for the assessment of reading skills. Lexical-decision tests and oral-reading tests were administered to 130 Dutch students from primary grades 1, 2, 3, 5, and 6. Correlations were moderate to high in low grades, but declined in the high grades. The reliability of the lexical-decision test assessed by means of a test–retest procedure was generally good. A second presentation of the lexical-decision test caused repetition effects (i.e., better performance on the second test), but generally remained within reasonable limits. The presence of different numbers of pseudowords (25% vs. 75%) in both lexical decision and oral reading, indicated that a large number of pseudowords made oral reading harder, but lexical decision easier. Educational and clinical implications are discussed.

Key words: Lexical decision, Literacy assessment, Oral reading, Test repetition, Word–pseudoword ratio

Introduction

In the Netherlands, reading skills of young students are usually assessed by means of standardized oral-reading tests. The most widely used tests require that students read aloud as quickly and accurately as possible, within a given time-span, a list of isolated words. All oral reading tests have an important requirement in common, namely, the production of an overt, oral response, a type of behavior that is usually not involved in fluent, competent reading. Kusters (1987) investigated this issue and showed that Dutch students who had been diagnosed as poor readers were better at meaning identification than at word pronunciation, suggesting that these poor readers often knew more about the meaning of words than was apparent from their oral-reading performance. A similar finding has been presented in English by Carlisle, Stone, and Katz (2001). They showed that both poor and good readers were better at lexical decision than oral reading, but the difference in performance between the two tasks was more pronounced in poor than in good readers. Lexical decision is probably the most widely used task in reading

research. In a lexical-decision task, readers are presented with strings of letters that either constitute a word or a pseudoword, and are asked to decide whether the presented stimulus forms a word or not. Pseudowords are defined as letter strings that are orthographically and phonologically legal in the relevant language. Usually, this task is performed individually on a computer. Accuracy and speed determine performance.

The lexical-decision task has been used as an alternative tool for the assessment of reading skill in young beginning readers by van Bon and his colleagues (van Bon, Bouwmans, Broeders, Hoevenaars & Jongeneelen, 2003; van Bon & Libert, 1997; van Bon, Tooren & van Eekelen, 2000). Van Bon and Libert (1997) presented young readers with a paper-and-pencil lexical-decision test. Sixty words mixed with 20 pseudowords were arranged in three columns on an A4-size piece of paper. Participants were asked to cross out as many pseudowords as possible in one minute. The number of correctly evaluated words and pseudowords determined performance. In the first study, van Bon and Libert found a distinct subtype of poor readers among the entire group of poor readers. These readers performed consistently worse on oral reading than on a comparable lexical-decision task, suggesting that particularly poor readers exhibit problems converting a word's phonology into speech.

In the second study, in which two groups of poor readers and a group of younger, reading-match students (grades 1 and 2) participated, van Bon et al. (2000) compared performance on a paper-and-pencil lexical-decision task with performance on two different oral-reading tests. One test constituted an oral-reading version equivalent to the lexical-decision test. The other test was the 'Een-minuut-test' [One-minute-test] of Brus and Voeten (1973); a standardized Dutch reading test, that involves reading aloud a list of single, isolated words of increasing complexity. The score on both oral-reading tests was the number of items read correctly in one minute. High correlations between lexical-decision and oral-reading performance emerged for both groups (r minimally 0.78), suggesting that lexical decision is a suitable alternative for oral reading.

In the third study, van Bon et al. (2003) again showed that lexical decision is a plausible tool to assess reading skills in readers in grades 2 and 3, albeit correlations between oral reading and lexical decision were lower in grade 3 than in grade 2. Moreover, correlations between the scores on the first and second administering of the test revealed an acceptable level of reliability (0.77 to 0.90) in all primary-grade levels. A final question they investigated was the effect of item structure: one-syllable (pseudo)words, two-syllable compound (pseudo)words, two-syllable non-compound (pseudo)words, and a mix of all three types of (pseudo)words. Correlations between scores on

lexical decision and oral reading were computed for all four lists. Results indicated that the item structure did not affect lexical-decision and oral-reading performance differentially.

The major theoretical issue here is whether oral reading and lexical decision are two valid, and to a certain extent interchangeable, tasks for the assessment of reading skills. To argue that one is a better measure than the other, suggests that the reading process is contextually independent, that is, that the word perception process is independent of task requirement. This, however, seems untenable (see for example, Bosman & de Groot, 1996; Van Orden, Holden, Podgornik & Aitchison, 1999). We believe that each task modulates the word identification process uniquely. Lexical decision and oral reading certainly share similarities with each other and with silent reading, but they are not identical. A number of studies show high correlations between the tasks, a finding that suggests that paper-and-pencil lexical decision is a possible alternative assessment procedure for oral reading. Moreover, lexical decision might be a useful alternative for students who suffer from speech problems; for example children with specific language impairments.

The present study aims at replicating and extending the findings presented in earlier work by van Bon and his colleagues. Primary-grade students (grades 1, 2, 3, 5 and 6) were presented with lists of items containing words and pseudowords. Half of the lists contained 25% pseudowords, whereas the other half had 75% pseudowords. The students were first asked to read silently through the lists and either cross out each word that does not exist (i.e., in lists that contained 25% pseudowords) or mark each extant word they encountered (i.e., in lists that contained 75% pseudowords). They were permitted one minute per list and the number of items evaluated correctly determined performance. Note, the type of item they had to attend to in the two tests is different. In the lexical-decision test with a low proportion of pseudowords, they were asked to mark the pseudowords, whereas in the same test with a high proportion of pseudowords, they had to mark the words. This way, they had to put pen to paper equally often in both conditions. Identical procedures with respect to required responses and with respect to the number of times the participant had to mark an item is impossible. Considering the age of the students, we chose to keep amount of motor movement, that is, number of times they had to put a mark, equal at the expense of slightly different types of responses. The same students were later presented with a second task, namely, reading aloud the same lists of items (both the 25% and the 75% pseudoword lists). They were again permitted one minute per list and the number of items read correctly determined performance on each list.

Four related issues will be investigated. First, the reliability of the lexical-decision test with 75% pseudowords needs to be established. In their earlier studies, van Bon and his colleagues investigated lexical-decision performance on a test with a low proportion of pseudowords (i.e., 25%) only. As in the study of van Bon et al (2003), we also expect high correlations between test and re-test performance in all grades.

A second research goal is the effect of repetition as a result of a second administering of the lexical-decision test with a high number of pseudowords. A large number of laboratory lexical-decision studies have shown repetition effects, usually referred to as repetition priming. Repetition priming is the improvement of speed or accuracy as a result of repeated exposures to a stimulus. When repetition priming is studied under masked and/or semantic priming conditions it usually only lasts a few seconds (Forster & Davis, 1984). There are, however, reports of so-called long-term repetition priming effects that last minutes (Bowers, Damian & Havelka, 2002; Oliphant, 1983), sometimes even days or months (Scarborough, Cortese & Scarborough, 1977; Sloman, Hayman, Ohta, Law & Tulving, 1988). Paper-and-pencil lexical decision, unlike laboratory lexical decision, does not allow testing performance at the level of the individual items. Thus, repetition effects in our tasks will have to be derived from performance at the level of the entire test. Although repetition effects are to be expected, strong repetition effects are undesirable with respect to multiple use in educational or clinical settings. Only weak effects are permissible, thus the size of the repetition effect needs to be established.

The third goal of the present study addresses the correlation between performance on oral reading and lexical decision in higher grades of primary school. Earlier studies focused on the lower grades and primarily at students with poor reading skills. In the present study, both lower and higher primary-grade students will participate. From a comparison between performance in lexical-decision and oral-reading tasks, we expect students to process more words in oral reading than in lexical decision, a common finding in English (Frost, Katz & Bentin, 1987; Katz & Feldman, 1983; Waters & Seidenberg, 1985) and Dutch (de Groot, 1985).

Findings related to the number of errors in naming tasks and lexical-decision tasks, reveal diverging results in the literature. Some authors report fewer errors in oral reading (Frost, Katz & Bentin, 1987; Seidenberg, Petersen, MacDonald & Plaut, 1996) than in lexical decision, whereas Katz and Feldman (1983) argue that the number of errors in pseudowords is higher in oral reading than in lexical decision, but comparable on words. We expect more errors in oral reading than in lexical decision, because oral reading requires correct pronunciation, which may cause supplementary errors.

The fourth and final main issue concerns the role of word–pseudoword ratio. Past and recent laboratory reading research proves the relevance of the word–pseudoword ratio variable. Taylor and Lupker (2001) showed that naming (i.e., oral reading) high-frequency words in a word-only condition was faster than in a condition in which the high-frequency words were mixed with pseudowords (see also, Lupker, Brown & Colombo, 1997; Monsell, Patterson, Graham, Hughes & Milroy, 1992). McQuade (1981) used a lexical-decision task and found longer 'no'-response latencies to pseudohomophones when they were embedded in a predominantly pseudohomophone condition than in a predominantly pseudoword-control condition, whereas this difference did not occur for the pseudoword controls (see for similar findings, Gibbs & Van Orden, 1998; Stone & Van Orden, 1993).

In all three studies, van Bon and colleagues used a low proportion of pseudowords (i.e., 25%) in their oral-reading and lexical-decision tasks in order to keep the motor component at a minimum. In the present study, two extreme (to maximize potential effects) conditions were applied: In one condition, the test contained a high proportion of pseudowords (75%), in the other condition a low proportion of pseudowords (25%) was present. With respect to performance differences in oral reading, we expect students to read fewer items and make more reading errors in the oral-reading test with 75% pseudowords than in the one with 25% pseudowords. Earlier findings show that pseudowords take longer to pronounce than words (e.g., Katz & Feldman, 1983; Lupker et al., 1997; Taylor & Lupker, 2001). With respect to lexical decision, we expect students to evaluate more items in one minute in the lexical-decision test with 25% pseudowords than in the lexical-decision test with 75% pseudowords, as a result of the so-called 'lexical-status effect'. The lexical-status effect refers to the finding that in a lexical-decision task positive judgments, that is, saying 'yes' to a word, are generally faster than negative judgments, that is, saying 'no' to a pseudoword (e.g., Stone & Van Orden, 1993; Gibbs & Van Orden, 1998; Underwood & Blatt, 1996).

Method

Participants

In this study, 130 students (54 boys, 76 girls) from one regular primary school in the Netherlands participated as subjects. They were recruited from grade 1 (n = 26, M age = 83 months), grade 2 (n = 29, M age = 95 months), grade 3 (n = 25, M age = 109 months), grade 5 (n = 28, M age = 132 months), and grade 6 (n = 22, M age = 147 months). One hundred twenty-three students (94.6%) were from the Netherlands and had Dutch as their native language;

the remaining seven students (5.4%) were originally from Germany, Surinam, Bosnia, and Iraq, and had Dutch as their second language. In the results' section, we show that performance of these students does not deviate from the native-Dutch students.

Materials

The materials in this study comprised 8 different lists. Four lists contained 20 words and 60 pseudowords. This test is referred to as the Word Identification Test (henceforth, WIT) because words had to be detected. The other four lists contained 60 words and 20 pseudowords. This test is referred to as the Pseudoword Identification Test (henceforth, PIT) because pseudowords had to be detected. One list of the PIT and WIT contained one-syllable CVCC- and CCVC-words (C stands for consonant and V for one or two vowels). Fifty-two items consisted of 4 letters, 28 items consisted of 5 letters. Word examples from this list are 'beest' [animal] and 'zalf' [salve] and pseudo-word examples are 'slin' and 'tars'. The second list of both the PIT and WIT contained two-syllable non-compound words. The length of the items ranged from 3 to 8 letters, with a mean of 5.7. Word examples are 'water' [water] and 'planten' [plants] and pseudoword examples are 'kasel' and 'kaspen'. The third list of the PIT and the WIT contained two-syllable compound words. The length of the items ranged from 6 to11 letters with a mean of 8.1 in the PIT and 7.1 in the WIT. Word examples are 'stoplicht' [traffic light] and 'verfpot' [paint-pot] and pseudoword examples are 'kleupstok' and 'schoorbeut'. The fourth list of both the PIT and the WIT contained words and pseudowords, including word types from all three categories mentioned above. The length of the items ranged from 4 to 11 letters with a mean of 6.1.

Almost all words were nouns, selected from the word familiarity ratings or frequency counts by Kohnstamm, Schaerlaekens, de Vries, Akkerhuis, and Froonincksx (1981), Staphorsius, Krom, and Geus (1988) and Praxis 14 (van der Geest, Swüste & Raeve, 1978). According to these sources, 7-year-old students are familiar with the selected words. By combining two arbitrary chosen items of a list in the PIT, we created both the pseudowords of the PIT and of the WIT. Two items (words and/or pseudowords) were combined and the initial, medial, or final consonant(cluster) was exchanged. For example, in List 1, the pseudoword 'ploor' was created by combining the words 'spoor' [trace] and 'ploeg' [plough]. In List 2, the pseudoword 'miffen' was created by combining the words 'midden' [middle] and 'stoffen [textiles]. In List 3, the syllables of two (compound) items were exchanged. Then, one letter was changed to make the new item unrecognizable as a word. For example, the pseudoword 'gandpoot' was created by combining the second part of

the word 'stoelpoot' [chairleg] with the first part of the word 'handdoek' [towel], and subsequently changing the h into g. Three requirements had to be met. First, the consonant-vowel structure and the number of letters of the pseudowords had to be the same as in the base words. Second, the spelling of the pseudowords had to be orthographically legal, keeping the pseudowords pronounceable. Third, no homophones or pseudohomophones were allowed.

Finally, four different item orders for each list were created. Order 1 was determined randomly. The remaining three orders were derived from the first one (the first word of order 1 is referred to as A, the middle words are referred to as M and N, and the last word is referred to as Z). The second order was the reversed order of the first one (items were ordered Z to A). The third order was from M to A, followed by Z to N. The fourth order was from N to Z, followed by A to M. The items of each list were printed in three columns on a single sheet of A4-paper. Two practice-tests, one for WIT and one for PIT, completed the set of materials. The items on these tests were CVC/VVC/VCV-words and pseudowords in identical proportions as in the experimental materials.

Procedure

Tasks. The study started with a class-administered session followed by an individual session. During the class-administered session, the students were asked to perform lexical decision on each of the eight lists. The different item orders of each list were randomly assigned to the students. It took about 15 minutes to complete all eight lists.

In each group, half of the students started with the LD-PIT, the other half started with the LD-WIT. The order of presentation of the lists within a test varied among students, such that each list was read by an equal number of students as first, second, third or fourth. Instruction and a practice test preceded each LD-test. First, in the instruction of both LD-tests, students were told that they would get lists with three columns of words. In both the LD-PIT and LD-WIT, the students were informed about the presence of pseudowords: In the LD-PIT, they were told that, besides a lot of words, they would encounter *some* nonsense words. In the LD-WIT, students were told that, besides some words, *a lot of* nonsense words were included. An explanation of 'nonsense words' was provided. Then, the students were asked to read the lists silently as quickly as possible and to cross out each pseudoword in the four lists of the LD-PIT and mark each word in case of all four lists of the WIT, working column by column. Self-corrections were permitted (the students were told how to perform the task). Finally, when they heard the digital alarm, the students were told that they had to stop reading immediately and put a line below the word (item) evaluated last. In both LD-tests,

the experimenter showed the procedure on blackboard with a few examples. After distribution of all lists, the experimenter summarized the instructions.

During the individual session, the child and the experimenter (a speech and language pathologist) sat in a separate room. Each child was asked to read aloud all eight lists. Recall, Phrasing? all lists are identical to the lists in the LD-test. Since this type of test is well known to the students, no practice test was included, but a short instruction preceded the OR-test to inform them about the presence of pseudowords (similarly, OR-WIT and OR-PIT). Subsequently students were asked to read aloud each list of items as fast and correctly as possible in one minute. The experimenter verified the correctness of each item and put a line below the item read last. Half of the group started with the OR-WIT and the other half started with the OR-PIT. The oral-reading test took up 10 to 15 minutes for each pupil.

Finally, four to six days after completion of the lexical-decision test, all students performed the LD-WIT a second time in order to determine reliability. This test is referred to as LD-WIT_{Rep}. To attenuate the possible influence of actual memory of particular items, the items in the LD-WIT_{Rep} were rearranged in order to break memory sets.

Scoring

Three performance variables on both lexical decision and oral reading were used in the analyses. Performance on the LD-test was determined by the number of items evaluated in one minute, the number of errors (false hits and missers), and the number of items evaluated correctly in one minute. False hits are wrongly crossed out words (PIT) or wrongly marked pseudowords (WIT). Missers are pseudowords (PIT) or words (WIT) mistakenly not marked. Performance on the OR-test was determined by the number of items read in one minute, the number of reading errors, and the number of items read correctly in one minute. Thus, the number of items evaluated correctly or the number of items read correctly is a reflection of both accuracy and speed.

An earlier study by van Bon et al. (2003) indicated that item structure did not differentially affect the findings. For that reason, we chose to collapse the scores of all four different lists (i.e., one-syllable word lists, two-syllable non-compound word lists, two-syllable compound word lists, and heterogeneous set of word lists) into one mean score per test. There is one exception, the reliability analysis of the LD-WIT was tested for all four lists separately, but in the remainder all test scores refer to the mean score of each test.

Results

Since seven students in our sample were native speakers of a language other than Dutch, we first established whether performance of these non-native speakers differed from their peers whose mother tongue was Dutch. A one-way ANOVA on the mean number of items processed correctly (either evaluated correctly in LD-tests or read correctly in OR-tests) of LD-PIT, LD-WIT, LD-WIT_{Rep}, OR-PIT, and OR-WIT showed that students whose native tongue was not Dutch did not differ significantly from students with Dutch as native language. Scores on the tests were: LD-PIT: 50.4 and 47.4; LD-WIT: 52.0 and 47.4; LD-WIT_{Rep} 54.8 and 52.1; OR-PIT: 49.9 and 51.8; OR-WIT: 37.8 and 38.9, respectively, all F-values < 1. It was, therefore, decided to include all students in subsequent analyses.

The result section starts with the presentation of the reliability analyses of the LD-WIT, followed by those of the test-repetition effect. Then, the analyses on the equivalence of the LD-tests to OR-tests will be presented, and finally we will discuss the effect of word–pseudoword ratio. Prior to the analyses presented below, four one-way ANOVA's were performed to test for list order effects. These analyses showed that none of the effects of order of presentation was significant, all F-values <1. Because of the absence of an order effect, this variable was removed from further analyses.

Reliability

A correlation analysis, for each grade separately, between the scores of the LD-WIT and the LD-WIT_{Rep} for all four lists (i.e., one-syllable word lists, two-syllable non-compound word lists, two-syllable compound word lists, and heterogeneous set of word lists) revealed that all test and retest scores correlated significantly, ranging from 0.55 to 0.91, except for two correlations in grade 6 (0.28 and 0.42). To evaluate a test's reliability for decisions at an individual level, we followed rules provided by Evers, van Vliet-Mulder, and Groot (2000). They state that reliability is considered insufficient if r is smaller than 0.70, sufficient if r is between 0.70 and 0.80 and reliability is good if r exceeds 0.80. According to these rules, reliability was good in grade 1 for all lists, sufficient in grade 2 for the first three lists, but insufficient for the list with a heterogeneous set of words ($r = 0.57$). The reliability in grades 3 and 5 was sufficient to good except for the heterogeneous list in grade 5 ($r = 0.60$). Finally in grade 6, the reliability was insufficient for the first three lists ($r = 0.55$, $r = 0.28$, and $r = 0.42$ respectively), but good for the heterogeneous list. The low correlations for the first list are possibly due to a ceiling effect.

Repetition

To test for repetition effects, performance on the LD-WIT (the test with 75% pseudowords) and LD-WIT$_{Rep}$ was compared. Table 1 presents the mean numbers of items evaluated, the mean numbers of errors, and the mean numbers of items evaluated correctly in both conditions. A five (grade: 1 vs. 2 vs. 3 vs. 5 vs. 6) by two (condition: LD-WIT vs. LD-WIT$_{Rep}$) ANOVA on all three performance variables with condition treated as within variable showed nearly identical outcomes for the number of items evaluated and the number of items evaluated correctly. Only the analysis on the number of items evaluated correctly will be discussed.

The main effect of grade was significant, $F(4,121) = 205.76$, $P < 0.0001$. Students in grade 1 evaluated significantly fewer items correctly than students in grade 2, who in turn evaluated significantly fewer items correctly than students in grade 3, who in turn evaluated significantly fewer items correctly than students in grades 5 and 6 (Fisher's *PLSD,* $P < 0.05$).

The main effect of condition was also significant, revealing a repetition effect. Performance on LD-WIT$_{Rep}$ was superior to LD-WIT, $F(1,121) = 30.03$, $P < 0.01$. The significant interaction between grade and condition, however, qualified this result, $F(4,121) = 2.63$, $P < 0.05$. Subsequent separate *t*-tests showed a significant repetition effect in all grades, except grade 1 (grade 2: $P < 0.01$; grade 3: $P < 0.01$; grade 5: $P < 0.01$; grade 6: $P < 0.05$). The same ANOVA on number of errors (including missers and false hits) revealed a significant main effect of grade only, $F(4,121) = 6.62$, $P < 0.01$. It appeared that grade 3 made significantly more errors than all other grades (Fisher's *PLSD,* $P < 0.05$). The mean number of errors in grades 1, 2, 5, and 6 were statistically similar. In all grades, the mean number of errors remained stable in the LD-WIT$_{Rep}$. The size of the repetition effects assessed in terms of increased percentage of items evaluated correctly were for grade 1: 6%, grade 2: 17%, grade 3: 7%, grade 5: 4%, and grade 6: 1%.

Correlations between LD and OR

To investigate the feasibility of lexical decision as an alternative for oral reading, Pearson's product moment correlations between LD-WIT and OR-WIT and between LD-PIT and OR-PIT were computed for all three performance variables. These results are presented in Table 2. In grades 1 and 2, lexical decision correlated strongly with oral reading regarding number of items processed (read or evaluated) and number of items processed correctly, both in the test with a high number of pseudowords (WIT) and in the one with a low number of pseudowords (PIT). In grades 3, 5, and 6, these correlations dropped considerably. With regard to errors, correlations were generally low

Table 1. Mean scores per minute and standard deviations in parentheses on the LD-WIT and LD-WIT$_{Rep}$ as a function of grade.

	Grade				
Tests	1	2	3	5	6
			Mean number of items evaluated		
LD-WIT	15.6 (8.0)	36.6 (12.3)	67.7 (12.6)	75.7 (7.5)	77.6 (3.0)
LD-WIT$_{Rep}$	16.5 (10.0)	41.6 (15.1)	71.5 (11.7)	78.6 (3.7)	78.4 (2.8)
Difference	0.9	5.0	3.8	2.9	0.8
			Mean number of errors		
LD-WIT	1.7 (1.0)	2.2 (1.6)	3.8 (3.0)	1.9 (1.2)	1.6 (1.2)
LD-WIT$_{Rep}$	1.7 (1.0)	2.1 (1.5)	3.4 (2.6)	2.0 (1.2)	1.3 (0.6)
Difference	0	−0.1	−0.4	1.0	−0.3
			Mean number of items evaluated correctly		
LD-WIT	13.9 (7.8)	34.4 (12.3)	63.9 (13.4)	73.8 (7.6)	75.9 (3.6)
LD-WIT$_{Rep}$	14.7 (9.6)	39.5 (14.8)	68.1 (12.5)	76.6 (4.0)	77.1 (3.0)
Difference	0.8	5.8	4.2	2.8	1.0

or even absent, which might be due to a relatively low number of errors on both tasks.

One further aspect of the equivalence of lexical decision to oral reading is the difference between number of items processed correctly in lexical decision and oral reading. Difference scores are presented in the lower part of Table 3. In all grades, the difference between OR-PIT and LD-PIT (tests with 25% pseudowords) did not significantly deviate from zero, overall $t(128) = 0.56$, $P = 0.58$. Stated differently, performance on the oral-reading test was identical to performance on the lexical-decision test when a low proportion of pseudowords was present in the tests.

The difference between OR-WIT and LD-WIT (tests with 75% pseudowords) did reach a significant level in all grades (all P's < 0.01). The value of the overall t-test was $t(127) = 12.4$, $p < 0.001$. All grades performed better in lexical decision than in oral reading when a high proportion of pseudowords was present in the tests.

To examine the cause of the lower number of items read correctly on OR-WIT, we compared the number of items read and the number of errors in the OR-WIT with the number of items evaluated and the number of errors (missers and false hits) in the LD-WIT. Separate t-tests for each grade revealed that all grades evaluated significantly more items in LD-WIT than

Table 2. Correlations between LD-tests and OR-tests for each grade.

Pearson's product-moment values	Grade 1	2	3	5	6
	Regarding number of items read/evaluated				
LD-WIT and OR-WIT	0.85	0.75	0.47	0.40	0.30
LD-PIT and OR-PIT	0.81	0.80	0.51	0.53	0.65
	Regarding number of reading/evaluation errors				
LD-WIT and OR-WIT	0.20	0.30	0.46	0.39	0.15
LD-PIT and OR-PIT	−0.04	0.80	0.64	0.51	0.54
	Regarding number of items read/evaluated correctly				
LD-WIT and OR-WIT	0.84	0.80	0.60	0.42	0.38
LD-PIT and OR-PIT	0.81	0.80	0.64	0.64	0.61

Note: Values exceeding 0.30 are significant at the 5% level.

in OR-WIT (P's < 0.01) except in grade 1 in which the difference was only marginally significant ($P < 0.10$). Separate analyses on errors showed that all grades made significantly more errors on the OR-WIT than the LD-WIT (all P's < 0.01). The absolute difference indicated that twice as many errors were made on OR-WIT than on LD-WIT, which will be comment on in detail in the discussion.

Word–pseudoword ratio

First, performance on OR-WIT was compared to performance on OR-PIT, revealing a significant overall effect regarding the mean number of items read correctly, $F(1,128) = 337.51$, $P < 0.001$. The mean number of items read correctly was higher on OR-PIT (M = 49.9, SD = 21.7) than on OR-WIT (M = 37.8, SD = 25.0). Separate t-tests for each grade showed that in all grades performance on OR-PIT was significantly better than performance on OR-WIT (all P's < 0.001). Mean and difference scores for each grade are summarized in Table 3.

To investigate the origin of this performance difference, the mean number of items read and the mean number of errors were analyzed. The overall analysis on the mean number of items read mimicked the results of the mean number of items read correctly $F(1,128) = 243.90$, $P < 0.01$. The mean number of items read was higher on OR-PIT (M = 52.9, SD = 24.6) than on OR-WIT (M = 42.6, SD = 21.3). Again, separate t-tests for each grade showed that in all grades, performance on OR-PIT was significantly better

Table 3. Mean number of items evaluated/read correctly, standard deviations, and difference scores on all tests as a function of grade.

Test		Grade 1	2	3	5	6
				Mean scores		
OR-PIT	Mean	12.8	40.5	60.2	68.8	70.9
	SD	10.8	15.1	15.8	9.6	9.2
OR-WIT	Mean	9.7	27.7	45.0	53.7	55.9
	SD	8.3	11.3	16.9	13.7	13.3
LD-PIT	Mean	12.4	37.0	60.6	71.6	73.7
	SD	7.4	12.8	11.7	7.2	6.8
LD-WIT	Mean	13.9	34.3	63.9	73.8	76.1
	SD	7.8	12.3	13.4	7.6	3.6
				Difference scores		
OR-PIT–LD-PIT		0.4	2.5	−0.4	−2.8	−2.8
OR-WIT–LD-WIT		−4.2	−6.6	−18.9	−20.1	−20.2
OR-PIT–OR-WIT		3.1	12.8	15.2	15.1	15.0
LD-PIT–LD-WIT		−1.5	2.7	−3.3	−2.2	−2.4

Note: Mean scores are based on the number of items read correctly on the OR-tests in one minute and on the number of items evaluated correctly on the LD-tests in one minute.

than performance on OR-WIT (all P's < 0.001). The overall analysis on the mean number of errors revealed a similar trend, $F(1,128) = 119.01$, $P < 0.01$. The mean number of errors was lower on the OR-PIT (M = 3.0, SD = 2.2) than on OR-WIT (M = 4.8, SD = 3.2), in all grades (all P's < 0.001). Note, correlations between OR-PIT and OR-WIT were high in all grades ($0.81 < r < 0.98$).

Second, performance on LD-WIT was compared to performance on LD-PIT, revealing a significant overall effect regarding the mean number of items evaluated correctly, $F(1,126) = 5.35$, $P < 0.05$. The mean number of items evaluated correctly was higher on LD-WIT (M = 51.6, SD = 26.2) than on LD-PIT (M = 50.3, SD = 25.3). Separate t-tests for each grade showed that this difference only reached significance in grades 1, 5, and 6 (P's < 0.05), but not in grades 2 and 3. Mean and difference scores for each grade are summarized in Table 3.

To investigate the origin of this performance difference, the mean number of items evaluated and the mean number of errors of the LD-WIT and LD-PIT were compared. The overall analysis on the mean number of items evaluated

did not reach significance, $F(1,126) = 2.16$, $P = 0.14$. The mean number of items evaluated on LD-PIT (M = 52.9, SD = 25.3) was not significantly different from LD-WIT (M = 53.8, SD = 26.3). Separate *t*-tests for each grade showed that in grades 1, 2, and 3 this difference was not significant either, but it reached significant levels in grades 5 and 6 (both P's < 0.05). More items were evaluated in the LD-WIT than in LD-PIT in these grades. The overall analysis on the mean number of errors (including missers and false hits) revealed a significant effect, $F(1,126) = 6.25$, $P < 0.01$. The mean number of errors was lower on LD-WIT (M = 2.2, SD = 1.9) than on LD-PIT (M = 2.6, SD = 2.1), but it only reached significance in grade 1 ($P < 0.01$). Note, however, that in all grades the effect was in the same direction and that correlations between LD-PIT and LD-WIT were high ($0.76 < r < 0.93$).

A final analysis concerned the distinction between false hits and missers on the lexical-decision task. In LD-WIT (high proportion of pseudowords), students had significantly more paper-and-pencil lexical decision 18 missers (failed to mark a word) than false hits (incorrectly marked a pseudoword), $F(1,127) = 9.30$, $P < 0.01$. In LD-PIT (low proportion of pseudowords), students also had significantly more missers (failed to mark a pseudoword) than false hits (incorrectly marked a word), $F(1,128) = 20.89$, $P < 0.001$. Although missing a target occurred more often than falsely identifying an item, the implications were different with regard to both tasks. In the condition in which they had to mark words, they made errors more often on words than on pseudowords, whereas in the condition in which they had to mark pseudowords, they made errors more often on pseudowords than on words.

Discussion

This study was designed to investigate four issues regarding paper-and-pencil lexical-decision tests. The first goal was to establish the reliability of a paper-and-pencil lexical-decision test with a large number of pseudowords. The second goal was to assess the extent to which test-repetition effects occur in a lexical-decision test with a high number of pseudowords. The third goal of the present study was to assess the correlation between paper-and-pencil lexical decision and oral reading in all grades of primary school. The fourth and final goal was to assess the effects of word-pseudoword ratio. Following the discussion of these four issues, we end our study with a discussion of the feasibility and merits of lexical decision as an alternative or supplementary tool for the assessment of reading skills.

The results of the reliability analysis indicated that for use in the educational or clinical setting, the lexical-decision test with a large number of

pseudowords (75%) appeared to be good in grade 1, sufficient in grades, 2, 3 and 5 for all but the heterogeneous list, whereas for grade 6 it was insufficient for all but the heterogeneous list. Earlier work by van Bon et al. (2003) also demonstrated somewhat lower reliability values in the higher grades (grade 3) than in the lower grades (grade 2). These findings suggest that a lexical-decision test with a large number of pseudowords can serve as, if not an alternative for, then at least as an appropriate supplementary tool to oral-reading tests, especially in the lower grades.

In all grades, except grade 1, a test-repetition effect emerged. This effect was mainly the result of an increased number of items evaluated correctly. The mean number of errors in the second administering of the test was equal to that in the first. In contrast to findings by Dannenbring and Briand (1982), who showed that participants responded faster as well as more accurately on repeated items, our participants were faster, but did not become more accurate nor did they become less accurate. Although reliable repetition effects emerged, the size of this effect varied in the different grades: Only in grade 2 an undesirable repetition effect emerged, a 17% performance increase. The reason for this relatively strong effect is unknown and subject of ongoing research. In all, it appears that paper-and-pencil lexical decision with a high proportion of pseudowords is suitable for reading assessment in most grades.

The analyses on the correlation between lexical decision and oral reading revealed that a paper-and-pencil lexical-decision test with a low number of pseudowords as well as one with a high number of pseudowords is a suitable alternative for oral reading in grades 1 and 2, but not in grades 3, 5, and 6. Although, correlations between oral reading and lexical decision in tests with a low number of pseudowords dropped substantially in the higher grades, performance in all grades on these tests was identical, as assessed by number of items processed correctly. However, with respect to the relation between oral reading and lexical decision in tests with a high number of pseudowords, performance in all grades was better on lexical decision than on oral reading, with higher grades showing larger differences. Performance differences were the result of more items evaluated as well as fewer errors in lexical decision than in oral reading.

An explanation for the decline in correlations between lexical decision and oral reading in the higher grades was provided by van Bon et al. (2000), who argued that to decide whether a string of letters forms a word or not, a reader needs to have orthographic knowledge. The acquisition of orthographic knowledge, however, requires reading experience. If young readers lack this information, they have to make the lexicality decision based on the phonological structure of the word only, a requirement similar to oral

reading. With increasing grade level students do acquire this orthographic knowledge, which enables them to rely on the orthographic structure as well as on the phonological aspect of the letter string. Thus, if similar procedures are used for lexical decision and oral reading in lower grades and different ones in higher grades, then this may account for relatively high correlations between lexical decision and oral reading in lower grades and low correlations in higher grades.

Ratcliff and McKoon (1988) provide yet another argument for the difference in task demands between lexical decision and oral reading. With regard to lexical decision, participants can provide a response based on some notion of familiarity such as "I have seen something like that before", and the response is binary. For example, some readers might recognize the word *Oxymoron* but may not know its meaning nor its proper pronunciation. In oral reading, knowledge of the meaning is not required either, but for proper pronunciation, the reader needs to know the word's phonology.

Word–pseudoword ratio caused differential effects in oral reading and lexical decision. Recall, for reasons explained in the introduction, in the lexical-decision test with a high proportion of pseudowords (LD-WIT), the reader had to mark the words, whereas in the same test with a low proportion of pseudowords (LD-PIT), the reader was asked to cross out the pseudowords. In all grades oral reading was harder than lexical decision when a high proportion of pseudowords was present, whereas oral reading was easier than lexical decision when a low proportion of pseudowords occurred in the test. Despite these performance differences, it appeared that correlations among tests regarding the number of items read/evaluated and the number of items read/evaluated correctly were high in grades 1 and 2 and most correlations were moderate in grades 3, 5, and 6. Correlations regarding the number of reading/evaluation errors were generally low.

Thus, it seems that a large number of pseudowords makes oral reading harder and lexical decision easier. The fact that a large number of pseudowords increases the difficulty of oral reading is not surprising. After all, reading aloud a word is easier than reading aloud a pseudoword. Frost et al. (1987), for instance, demonstrated that in English, words were named significantly more slowly in an 80%-pseudoword condition than in a 20%-pseudoword condition. But, why is lexical decision easier when the number of pseudowords increases, a finding which contradicts the expected lexical-status effect? Recall, in our study, the number of pseudowords as well as the task demands differed for lexical decision and oral reading. Therefore, a comparison to findings of other studies might be difficult or impossible. The question is which factors affect deciding whether an item is a word or a pseudoword? In lexical decision, a reader may apply specific criteria

for responding. It is suggested that these criteria are affected by factors as relative discriminability of the stimuli, and biases and expectations of the participants (Waters & Seidenberg, 1985). Discriminability between items is easier when more pseudowords are included and the required response is marking the words. The high-frequency words present in our materials and embedded in many unknown orthographic and phonological structures may immediately leap to the eye (see for more detailed information Seidenberg & McClelland, 1989). At the same time, it is also possible that the higher proportion of pseudowords lowers the threshold for accepting an item as nonexistent. Moreover, when young readers were asked to mark the words, they made more errors on words than on pseudowords. When they were required to cross out the pseudowords, they made more errors on pseudowords than on words. These results indicate a bias of the reader. In a large quantity of words, the reader is expecting words rather than pseudowords, but in a large quantity of pseudowords, the reader expects pseudowords rather than words.

In sum, the results of the reliability and repetition analyses suggest that paper-and-pencil lexical decision with a high number of pseudowords is a reliable measure for the assessment of reading skills in grades 1, 3, and 5. The results of the correlation analysis, however, indicates that lexical decision both with a low and with a high number of pseudowords are valid alternative tools for oral reading in lower grades only (i.e., grades 1 and 2). A practical implication of this finding is that tests with a low proportion of words may serve as a useful alternative in case students have small vocabularies. The results of the word–pseudoword ratio analyses revealing that in all grades a large number of pseudowords made oral reading harder and lexical decision easier, also indicate that lexical decision is not merely a simple substitute for oral reading.

Thus, paper-and-pencil lexical decision is a reliable tool, with only small repetition effects, but it appears to measure a slightly different aspect of reading skills than oral reading does. Does this conclusion force us to reject lexical decision as a suitable tool for oral reading? We believe, the answer is no. Although oral reading has been the standard tool for the assessment of reading skills for good reasons, it does not necessarily have primacy. Oral reading involves knowledge of letter-sound relations and proper pronunciation of the written word. The fact that words are named faster than pseudowords indicates that knowledge of word meanings plays also a significant role in word perception. However, the fact that children and adults are capable of reading pseudowords aloud indicates that meaning activation is not obligatory. Lexical decision also requires knowledge of letter-sound relations but in a lexical-decision test, minimal meaning activation is necessary, in the sense

that one has to know at least whether the item is an existing word or not. The exact meaning of the word is not required. Experiments by Van Orden et al. (1992), in which subjects have to make lexical decisions on pseudohomophones point out that more errors are made on pseudohomophones than on non-homophonic pseudowords. Bosman and de Groot (1996) report the same findings for lexical decision in children. Chumbley and Balota (1984) also argue that knowledge of the meaning of a word may affect lexical-decision performance. In short, lexical decision most likely involves minimal knowledge of word meanings but it does not require proper pronunciation of the letter string. In other words, oral reading and lexical decision appear to share one major aspect in common, namely, knowledge of letter-sound relations and to a lesser extent meaning activation. In experienced readers overt pronunciation is generally absent, they read silently, but in order to extract information from a text, they need to have word knowledge. In other words, lexical decision seems to share more characteristics with silent reading than oral reading does. This, however, does not imply that lexical decision should become the primary measure for assessing reading skills. After all, oral-reading tests have proven their usefulness and validity over the last decades, but there are, given the findings of the present study, two reasons for future investigation into lexical decision as an alternative, rather than a substitute for oral reading. The first reason was discussed in the introduction. Substantial groups of students with special needs (e.g., students with specific language impairments, students with hearing impairments and deaf students) have problems with the production of an overt response. The second reason involves the relationship between lexical decision and reading comprehension. Given the aspect of word knowledge in lexical decision it may be worthwhile to investigate whether lexical decision has a perhaps even better predictive value with respect to reading comprehension.

Acknowledgements

We are greatly indebted to the children and the teachers of "Basisschool St. Martinus", a regular primary school in Millingen (The Netherlands), for their participation in our study. The recommendations and insightful comments of three anonymous reviewers are greatly appreciated.

References

Bosman, A.M.T. & de Groot, A.M.B. (1996). Phonologic mediation is fundamental to reading: Evidence from beginning readers. *The Quarterly Journal of Experimental Psychology*, *49A*, 715–744.

Bowers, J.S., Damian, M.F. & Havelka, J. (2002). Can distributed orthographic knowledge support word-specific long-term priming? Apparently so. *Journal of Memory and Language*, *46*, 24–38.

Brus, B.T. & Voeten, M.J.M. (1973). *Een-Minuut-Test* [*One-Minute test*]. Nijmegen, The Netherlands: Berkhout.

Carlisle, J.F., Stone, C.A. & Katz, L.A. (2001). The effects of phonological transparency on reading derived words. *Annals of Dyslexia*, *51*, 249–274.

Chumbley, J.I. & Balota, D.A. (1984). A word's meaning affects the decision in lexical decision. *Memory and Cognition*, *12*, 590–606.

Dannenbring, G.L. & Briand, K. (1982). Semantic priming and the word repetition effect in a lexical decision task. *Canadian Journal of Psychology*, *36*, 435–444.

de Groot, A.M.B. (1985). Word-context effects in word naming and lexical decision. *The Quarterly Journal of Experimental Psychology*, *37*, 281–297.

Evers, A., van Vliet-Mulder, J.C. & Groot, C.J. (2000). *Documentatie van tests en testresearch in Nederland* [*Documentation of tests and test research in the Netherlands*]. Assen, The Netherlands: Van Gorcum.

Frost, R., Katz, L. & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 104–115.

Forster, K.I. & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 680–698.

Gibbs, P. & Van Orden, G.C. (1998). Pathway selection's utility for control of word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1162–1187.

Katz, L. & Feldman, L.B. (1983). Relation between pronunciation and recognition of printed words in deep and shallow orthographies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 157–166.

Kohnstamm, G.A., Schaerlaekens, A.M., de Vries, A.K., Akkerhuis, G.W. & Froonincksx, M. (1981). *Nieuwe streeflijst woordenschat voor 6-jarigen* [*Target list vocabulary for 6-year-old children*]. Lisse, The Netherlands: Swets & Zeitlinger.

Kusters, E.D.M. (1987). Self-corrections in oral reading: Some aspects of the reading process of good and poor readers. Unpublished doctoral dissertation. University of Nijmegen, Nijmegen, the Netherlands.

Lupker, S.J., Brown, P. & Colombo L. (1997). Strategic control in a naming task: changing routes or changing deadlines? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 570–590.

McQuade, D. (1981). Variable reliance on phonological information in visual word recognition. *Language and Speech*, *24*, 99–109.

Monsell, S., Patterson, K., Graham, A., Hughes, C.H. & Milroy, R. (1992). Lexical and sublexical translations of spelling to sound: Strategic anticipation of lexical status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 452–467.

Oliphant, G.W. (1983). Repetition and recency effects in word recognition. *Australian Journal of Psychology*, *35*, 393–403.

Ratcliff, R. & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*, 385–408.

Scarborough, D.L., Cortese, C. & Scarborough, H.S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 1–17.

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

Seidenberg, M.S., Petersen, A., MacDonald, M.C. & Plaut, D.C. (1996). Pseudohomophone effects and models of word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 48–62.

Sloman, S.A., Hayman, C.A.G., Ohta, H., Law, J. & Tulving, E. (1988). Forgetting in primed fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 223–239.

Staphorsius, K., Krom, R.S.H. & de Geus, K. (1988). *Frequenties van woordvormen en letterposities in jeugdlectuur* [*Frequencies of word forms and letter positions in youth literature*]. Arnhem, The Netherlands: CITO.

Stone, G.O. & Van Orden, G.C. (1993). Strategic control of processing in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 744–774.

Taylor, T.E. & Lupker, S.J. (2001). Sequential effects in naming: A time-criterion account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 117–138.

Underwood, G. & Blatt, V. (1996). *Reading and understanding*. Oxford: Blackwell Publishers Ltd.

van Bon, W.H.J., Bouwmans, M., Broeders, I., Hoevenaars, L. & Jongeneelen, J.E. (2003). Een klassikale toets voor 'technische leervaardigheid': vragen van validiteit en betrouwbaarheid [Lexical decision and oral reading: validity and reliability]. *Tijdschrift voor Orthopedagogiek*, *42*, 71–86.

van Bon, W.H.J. & Libert, J.E.A. (1997). Oral reading and silent reading compared: Evidence for a subtype of poor readers. *Polish Psychological Bulletin*, *28*, 59–70.

van Bon, W.H.J., Tooren, P.H. & van Eekelen, K.W.J.M. (2000). Lexical decision and oral reading by poor and normal readers. *European Journal of Psychology of Education*, *3*, 259–270.

van der Geest, A, Swüste, W. & Raeve, J. (1978). *Praxis 14: Spellingwijzer* [*Praxis 14: Spelling guide*]. Hertogenbosch, The Netherlands: Malmberg.

Van Orden, G.C., Holden, J.G., Podgornik, M.N. & Aitchison, C.S. (1999). What swimming says about reading: Coordination, context, and homophone errors. *Ecological psychology*, *11*, 45–79.

Van Orden, G.C., Stone, G.O., Garlington, K.L., Markson, L.R., Pinnt, G.S., Simonfy, C.M. & Brichetto, T. (1992). "Assembled" phonology and reading: A case study in how theoretical perspective shapes empirical investigation. In R. Frost and L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 249–292). Amsterdam: Elsevier Science Publishers B.V.

Waters, G.S. & Seidenberg, M.S. (1985). Spelling-sound effects in reading: Time-course and decision criteria. *Memory and Cognition*, *13*, 557–572.

Address for correspondence: Martine A.R. Gijsel, Faculty of Social Sciences, Department of Special Education, University of Nijmegen, PO Box 9104, 6500 HE Nijmegen, The Netherlands
E-mail: m.gijsel@ped.kun.nl