Learning and Individual Differences 56 (2017) 159–166

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Development of a word reading test: Identifying students at-risk for reading problems

Séli Chaves-Sousa [a], Sandra Santos [a,*], Fernanda Leopoldina Viana [b], Ana Paula Vale [c], Irene Cadime [a], Gerardo Prieto [d], Iolanda Ribeiro [a]

[a] School of Psychology, University of Minho, Portugal; Campus de Gualtar, 4710-057 Braga, Portugal
[b] Institute of Education, University of Minho, Portugal; Campus de Gualtar, 4710-057 Braga, Portugal
[c] Education and Psychology Department, University of Trás-os-Montes e Alto Douro, Portugal; ECHS, DEP- CIFOP - Rua Dr. Manuel Cardona, 5000-558 Vila Real, Portugal
[d] Faculty of Psychology, University of Salamanca, Spain; Campus Ciudad Jardín. Avda. de la Merced 109-131, 37005 Salamanca, Spain

ARTICLE INFO

Article history:
Received 28 September 2015
Received in revised form 15 September 2016
Accepted 20 November 2016

Keywords:
Rasch model
Word reading
Assessment
At-risk readers
Reading problems

ABSTRACT

The aim of this study was twofold. In Study 1, we described the development of four forms of a test of word reading (TLP – Teste de Leitura de Palavras) for elementary school children (grades 1 to 4), using the Rasch model. An initial pool of 142 words was selected and tested on 905 Portuguese students. Rasch analyses allowed the development of a shorter version of the test for each grade with adjusted values concerning reliability coefficients and item local independence. In Study 2 ($n = 325$), the classification accuracy of the TLP to identify at-risk students for reading problems was examined based on several indices. Results indicated that each test form of the TLP presented overall satisfactory classification accuracy in identifying at-risk readers with a criterion of 0.80 to set the sensitivity levels.

1. Introduction

The importance of word recognition, defined as "the ability to read printed words without the aid of context" (Adlof, Catts, & Little, 2006), is well documented in the literature (see Goff, Pratt, & Ong, 2005, for a review). Previous research has demonstrated the relationship between word recognition and reading comprehension in students from different grades (from grade 1 to grade 9), with positive correlations varying between moderate (in later grades) and high (in early grades) (e.g., Francis, Fletcher, Catts, & Tomblin, 2005; Gough, Hoover, & Peterson, 1996). Moreover, developmental differences in this relationship are observed, as expressed by the difference in the percentage of explained variance across grades, with the contribution of word recognition skills for reading comprehension declining across grades (Adlof et al., 2006; Catts, Hogan, & Adlof, 2005). Because accurate word recognition is a necessary condition for reading fluency, poor word reading accuracy would negatively affect reading fluency (Hudson, Lane, & Pullen, 2005). The contribution of word recognition to reading fluency has been supported by several studies that found a high correlation between word recognition and reading fluency in children from the second to the eighth grades (e.g., Adlof et al., 2006; Meisinger, Bloom, & Hynd, 2010).

The lack of automaticity in word recognition highly compromises reading performance and is one of the main sources of reading disabilities (Høien-Tengesdal & Tønnessen, 2011; Lewandowski, Begeny, & Rogers, 2006). Several studies found that disabled readers of several grades present difficulties in word recognition (Daane, Campbell, Grigg, Goodman, & Oranje, 2005; Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003; Rack, Snowling, & Olson, 1992). Word recognition is a good predictor of reading comprehension performance in students at the beginning of reading acquisition and in students with reading disabilities (Vellutino, Fletcher, Snowling, & Scanlon, 2004). Fuchs, Fuchs, and Compton (2004) found, in a sample of 151 at-risk readers in grade 1, that word and pseudo-word recognition measured in the fall were both predictive of reading comprehension measured in the spring. The results from Berninger, Abbott, Vermeulen, and Fulton (2006) also showed that word reading accuracy uniquely contributed to reading comprehension in at-risk readers in grade 2.

The accurate identification of disabled readers needs to occur as soon as possible to reduce the incidence or the severity of reading problems (Jenkins & O'Connor, 2002). Screening measures are used to predict an outcome (i.e., the criterion measure) months or years in advance and allow the quick identification of students who might be at-risk for reading problems (Johnson, Pool & Carter, n.d.). When using screening measures we expect to accurately identify students who are at-risk (true positives that fail the screen and the criterion

* Corresponding author at: Escola de Psicologia, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal.
E-mail address: sandra.css@gmail.com (S. Santos).

http://dx.doi.org/10.1016/j.lindif.2016.11.008
1041-6080/

measure) or who are not at-risk (true negatives that pass the screen and the criterion measure). However, it is also predictable that some inaccuracy in this identification could happen, so, a number of false positives (i.e., fail the screen but pass the criterion measure) and false negatives (i.e., pass the screen but fail the criterion measure) is expected to occur. The risk criterion to classify students' reading performance on the outcome measure is not consensual among researchers: some authors consider one standard deviation (SD) below the mean, whereas others draw the line at the 25th or 30th percentile (Jenkins, Hudson, & Johnson, 2007, for a review).

Classification accuracy studies provide evidence of screening measures' validity (Jenkins et al., 2007) through the analysis of the screening measure's accuracy of classification (i.e., the percentage of students correctly classified as either true positives or true negatives) (Johnson, Jenkins, Petscher, & Catts, 2009) and the use of two statistics: sensitivity (i.e., "the percentage of the individuals classified as unsuccessful on the future criterion reading test who were correctly identified as at-risk on the screen", Johnson et al., 2009, p. 175) and specificity (i.e., "the percentage of the individuals classified as successful on the future criterion reading test who were correctly identified as not at-risk on the screen", Johnson et al., 2009, p. 175). As sensitivity levels increase, the number of false negatives decreases, whereas as specificity levels increase, the number of false positives decreases. This misclassification has costs: false negatives will not receive the support they need and the false positives will be assigned to unneeded educational support, wasting human and financial resources necessary for intervention with students with reading disabilities (Compton, Fuchs, Fuchs, & Bryant, 2006; Stevens, 1992). Due to the consequences of flagging false negatives (Klingbeil, McComas, Burns, & Helman, 2015; Slocum, 2002), most of the test developers choose to maximize sensitivity, which reduces the number of false negatives (Johnson et al., 2009). Concerning the minimally acceptable levels of sensitivity and specificity, some authors recommended 0.90 for sensitivity and 0.80 for specificity (Compton et al., 2006) whereas others suggested a value for sensitivity of 0.80 and specificity of 0.70 (Johnson et al., 2009).

Several authors recommended the development of age-based reading screening measures with an appropriate level of difficulty and that are sensitive to the different levels of reading development to warrant good classification accuracy (Jenkins et al., 2007; Jenkins & Johnson, n.d.). The reading skill that screens should target to accurately identify at-risk readers may be different depending on the reading skill that is pertinent to the grade in which the screening measure will be administered. In the first grade, screens should target a word identification fluency measure; in grade 2, they should include, beyond the word identification fluency, the oral reading fluency (Jenkins et al., 2007; Jenkins & Johnson, n.d.). Beyond this grade, the identification of at-risk readers has been mainly done by reading fluency screening measures (Jenkins et al., 2007). However, at the 4th grade some students still have difficulties in word recognition (Lipka, Lesaux, & Siegel, 2006), therefore it is necessary to examine to what extent word recognition tests accurately identify at risk-readers at grades 3 and 4.

In most word recognition measures (e.g., the Word Identification and Word Attack subtests of the Woodcock-Reading Mastery Test - WRMT-III, Woodcock, 2011; and the Reading subtest of the Wide Range Achievement Test - WRAT-4, Wilkinson & Robertson, 2006) children are asked to read aloud a set of words, usually organized by increasing difficulty, which may be presented in lists or isolated, with no time limit. In the Portugal, only three word reading tests are available. The Word Recognition Test (PRP - Prova de Reconhecimento de Palavras, Viana, Ribeiro, Maia, & Santos, 2013) is a screening test composed of 40 regular and frequent words that assesses silent word recognition in students from first to fourth grade. The word reading subtest of the ALEPE (Sucena & Castro, 2012) is composed of three lists of words for students from grades 2 to 4. Items from all of the lists include regular words, irregular words and rule-based words. Ceiling effects in the word recognition subtest for grades 2 to 4 were observed. The Portuguese adaptation of the Psycholinguistic Assessments of Language Processing in Aphasia (Castro et al., 2007) includes word reading tasks to assess word reading accuracy in individuals with aphasia and persons who may present language problems. However, no information concerning psychometric properties is provided.

The major limitation of the Portuguese word recognition measures is their low ability to distinguish word reading skills of students from different grades, with ceiling effects that make it more difficult to accurately identify students who are at-risk for reading problems. Word reading tests should include items with different levels of difficulty to allow the accurate assessment of children with different word reading skills. One solution is the development of a different word reading list for each elementary grade by selecting the more appropriate words to distinguish readers from each grade. This process can be achieved using the Rasch model, an Item Response Theory model, that provides an estimation of the item difficulty and the person's ability. Thus, using the Rasch analysis to evaluate the appropriateness of word difficulty to assess a particular group might be useful in the development of such word lists.

2. Study purposes

This work is divided into two studies. The aim of study 1 was to develop, for European Portuguese, a word reading screening test (TLP-Teste de Leitura de Palavras) that surpasses the limitations of the word reading tests available, with four test forms, one for each elementary grade (grades 1 to 4) using Rasch model analyses. Due to the lack of studies that investigate the classification accuracy of word reading screening measures beyond grade 2, the purpose of study 2 was to investigate the classification accuracy of the TLP as a screening measure to identify at-risk readers across elementary school grades (grades 1 to 4). To pursue this goal, we analysed: (a) the power of the TLP test forms to predict oral reading fluency problems, as well as reading comprehension problems; and (b) the accuracy classification indices of the TLP in identifying at-risk readers in the first, second, third and fourth grades.

3. Study 1

Study 1 describes the development of the TLP for students from grades 1 to 4, with one form per grade, as well as their psychometric properties.

3.1. Method

3.1.1. Participants

The participants in study 1 consisted of 905 Portuguese students between 6 and 11 years old (Mean = 8.01, SD = 1.23). Two hundred and six first-graders (50.5% male), 229 second-graders (47.2% male), 235 third-graders (47.2% male) and 235 fourth-graders (46.4% male) participated in this study. The students attended both private (18.1%) and public (81.9%) schools located in the northern and central regions of Portugal. National data indicated that in 2013, approximately 11.7% of elementary students attended private schools. The students were all native European Portuguese speakers. Bilingual children and students with educational special needs that severely affect learning were not included in the study.

3.1.2. Measure and procedure

A pool of 142 Portuguese words was used to develop the four forms of the TLP. The pool was composed of words varying in: frequency (high vs. low), regularity (regular vs. irregular) and length (short vs. long). The word frequency was based on information available in European Portuguese lexical databases (Bacelar do Nascimento, Garcia Marques, & Segura da Cruz, 1987; Bacelar do Nascimento, Rivenc, & Segura da Cruz, 1987; Gomes & Castro, 2003). The regularity was controlled taking into account the irregular grapheme-phoneme conversion and the

word accent. Words with more than two syllables were classified as long words. The pool contained 75 high frequency and 67 low frequency words, 72 short and 70 long words and also 100 regular and 42 irregular words. European Portuguese is classified as a regular orthography in reading and as an intermediate depth orthography in writing (Seymour, Aro, & Erskine, 2003); consequently, the frequency of irregular words is smaller when compared to the regular ones. Words were also selected considering the most frequent syllabic structure in European Portuguese (cf. Appendix). Monosyllabic words were not included because they are not frequent in the Portuguese language (Lopes, 2011).

The pool of 142 words was administered to students individually. Children were instructed to read aloud the words, which were displayed one at a time on a computer. The words' display order was randomized and then, items were presented in the same order to all students. The testing procedure included 10 example items. The administration of the 142 words was split into two blocks of 47 words and a third block of 48 words. Two minutes of break time were given to the children between each block. There was no time limit for word presentation on the monitor or to perform the test. Errors (omissions, insertions, mispronunciations and substitutions) were scored with 0, and correct responses (including regionalisms and self-corrections) were scored with 1. The total score corresponds to the number of words read correctly.

The data collection was authorized by the Portuguese Ministry of Education, school boards and parents. The assessments were conducted by trained psychologists in a separate room in the students' schools. Each assessment session lasted from 15 to 40 min.

3.1.3. Data analyses

The test development followed several steps: (a) the psychometric analysis of an initial pool of items using the Rasch model, (b) the item selection for each test form, and (c) the analysis of the reliability of the four test forms.

Before the development of the test forms, the prerequisite of unidimensionality of the pool of items to use Rasch analyses was tested with a confirmatory factor analysis (CFA), using the WLSMV estimator. The overall model's goodness of fit was evaluated according to the following fit indices: the Chi-Square (χ^2) test (a non-significant p value from the chi-square test indicates a good fit, Byrne, 2012), the Root Mean Square Error of Approximation (RMSEA) (values <0.05 indicates a good fit, Browne & Cudeck, 1993), the Comparative Fit Index (CFI) and the Tucker–Lewis Index (TLI). CFI and TLI values higher than 0.95 are considered to indicate a good fit (Hu & Bentler, 1999). The analyses were performed using the M*plus* software version 6.1 (Muthén & Muthén, 2010).

Rasch model analyses were performed in three phases using WINSTEPS software, version 3.72.0 (Linacre, 2011b). First, the initial pool of 142 items was calibrated separately for each grade. Person ability and item difficulty parameters were estimated. The fit of the data to the model was assessed by calculating infit and outfit mean square indices. According to Linacre (2002), infit and outfit values higher than 2.0 indicate misfit in the data. The assumption of local independence was tested using a principal component analysis of residuals. The residual correlations should be smaller than 0.70 to support the local independence assumption (Linacre, 2011a). The reliability was studied by computing person- and item-separation reliability coefficients (PSR and ISR), as well as the Kuder-Richardson formula 20 (KR-20). Values higher than 0.70 were considered acceptable (Nunnally, 1978). Second, the words for each test form were selected according to the item's analysis results. Items that revealed a) misfit indices, b) point measure correlations lower than 0.20 or c) residual correlations higher than 0.70 were excluded. Finally, the reliability coefficients were re-calculated.

3.2. Results

3.2.1. Unidimensionality

The one-dimensional structure for the initial pool of items was supported by the CFA with the excellent fit of the one factor model, χ^2 (9869) = 11.077.69, $p < 0.001$, CFI = 0.99, TLI = 0.98, RMSEA = 0.01.

3.2.2. Item analysis

Table 1 presents the descriptive statistics and infit and outfit indices for item difficulty and person ability by grade. In grade 1, the highest ability value was 3.34, and the most difficult word presented a difficulty value of 4.39. These data suggested that there were several items that were too difficult for first graders. None of the infit indices were above the value of 2.0. Five items had outfit indices higher than 2.0. The point-measure correlations were positive and varied between 0.15 and 0.69; only six items presented point-measure correlations lower than 0.20. The percentage of students with outfit higher than 2.0 was 3.40% ($n = 7$).

In grade 2, item difficulty values for the easiest and for the most difficult items were within the minimum and maximum limits for person ability values (see Table 1). None of the 142 words was determined to be too difficult or too easy for second graders to perform. The infit indices were smaller than 2.0. Five items showed outfit indices above 2.0. Only one item presented a point-measure correlation lower than 0.20. The percentage of students with outfit values higher than 2.0 was 3.06% ($n = 7$).

In grade 3, 40 items presented difficulty levels lower than the minimum person ability value, indicating the high level of ease of these words for third graders. No item or students had an infit value higher than 2.0. Three items had an outfit value higher than 2.0. Fifteen items showed point-measure correlations lower than 0.20. The percentage of subjects with outfit higher than 2.0 was small (2.98%, $n = 7$).

Regarding grade 4, 54 items were excessively easy for this grade group. None of the items had infit values above 2.0. Five items presented outfit indices above 2.0. Twenty-seven words presented point-measure correlations lower than 0.20. The percentage of students with outfit values higher than 2.0 was small (4.68%, $n = 11$).

Only three pairs of items in grade 4 presented residual correlations higher than 0.70. All of the remaining items obtained residual

Table 1
Descriptive statistics for estimated parameters by grade.

Grade	Parameters	Mean	SD	Min	Max
1	Item difficulty	0.00	1.44	−3.62	4.39
	Person ability	0.06	1.57	−5.05	3.34
	Person infit	0.99	0.16	0.69	1.98
	Person outfit	1.03	0.44	0.24	4.51
	Item infit	1.00	0.18	0.70	1.63
	Item outfit	1.03	0.42	0.31	2.59
2	Item difficulty	0.00	1.55	−3.73	3.82
	Person ability	1.39	1.44	−4.79	4.62
	Person infit	1.00	0.15	0.60	1.61
	Person outfit	0.98	0.42	0.15	3.15
	Item infit	0.99	0.18	0.61	1.61
	Item outfit	0.98	0.46	0.35	3.57
3	Item difficulty	0.00	1.72	−3.84	4.00
	Person ability	2.33	1.17	−0.97	6.23
	Person infit	1.01	0.13	0.72	1.42
	Person outfit	0.92	0.67	0.24	7.31
	Item infit	0.99	0.13	0.69	1.48
	Item outfit	0.92	0.60	0.10	6.06
4	Item difficulty	0.00	1.72	−3.28	3.80
	Person ability	2.86	1.14	−0.86	6.18
	Person infit	1.00	0.12	0.72	1.33
	Person outfit	0.90	0.74	0.19	8.33
	Item infit	1.00	0.10	0.75	1.42
	Item outfit	0.90	0.40	0.14	2.27

Note. SD = standard deviation; Min = minimum value; Max = maximum value.

correlations smaller than 0.70, supporting the assumption of local independence (Linacre, 2011a).

The ISR, the PSR and the KR-20 were very high in each grade. The ISR values were 0.98, 0.98, 0.97 and 0.95 in grades 1, 2, 3, and 4, respectively. The PSR values were 0.98, 0.97, 0.93 and 0.91 in grades 1, 2, 3, and 4, respectively. Finally, the KR-20 values were 0.98, 0.98, 0.95, and 0.94 in grades 1, 2, 3, and 4, respectively.

3.2.3. Item selection and development of the test forms

Following the previous criteria for item selection, 30 items were selected for each test form. In the test forms, the number of low-frequency words increases as the school grade increases from grades 1 to 4, and the opposite trend occurs with high-frequency words: the number of regular and irregular words is similar between the TLP-1 (21 regular and 9 irregular words) and the TLP-2 (22 regular and 8 irregular words); the TLP-3 and the TLP-4 present a higher number of irregular words (16 and 14, respectively) than the TLP-1 and the TLP-2.

3.2.4. Reliability analyses

The reliability analyses were performed for each group of items that constitute the final test forms of the TLP. The person-separation reliability, KR-20 and item-separation reliability coefficients for each test form were high to very high (see Table 2).

4. Study 2

The aims of study 2 were: to examine the ability of each test form of the TLP to predict the at-risk status on reading fluency and reading comprehension; to analyse the classification accuracy indices of each test form to identify at-risk students for reading fluency or reading comprehension problems.

4.1. Method

4.1.1. Participants

Participants in study 2 ($n = 325$) consisted of 86 first-graders (58.1% male), 79 second-graders (51.9% male), 94 third-graders (55.3% male) and 66 fourth-graders (53.0% male) from six public schools from the North of Portugal with ages from 6 to 10 years old (Mean = 7.91, SD = 1.21). The students' selection criteria in this study are similar to those used in study 1.

4.1.2. Measures and procedure

The TLP-1, TLP-2, TLP-3 and TLP-4 were administered to the participants in this study.

The Reading Comprehension Test (TCL – Teste da Compreensão da Leitura; Cadime et al., 2013, Cadime, Ribeiro, Viana, Santos, & Prieto, 2014). The TCL is a norm-referenced test that assesses four components of reading comprehension (literal comprehension, inferential comprehension, reorganization, and critical comprehension) in students from second to fourth grades through three forms (one for each grade). It can be administrated individually or in group. The text is identical across the test forms and consists of poems, as well as narrative, informative, and instructional sequences. Each form has 30 multiple-choice items with four options, only one of which is correct. The TCL was developed with Rasch analyses that allowed the item selection based on the person ability and the item difficulty. Therefore, items too easy and with local dependence were excluded. Confirmatory factor analysis supported a one-factor structure. Correlation coefficients with other reading tests were low to moderate and statistically significant. The reliability coefficients ranged between 0.70 and 0.98.

The Test of Reading Fluency (TFL – Teste de Fluência da Leitura, Ribeiro et al., 2014). The TFL assesses oral reading fluency (i.e., the number of words read correctly per minute) in European Portuguese students from grades 1 to 4. The test application is individual. Children from grades 1 to 4 were asked to read aloud the same text with 467 words (approximately 46% of the words were low frequency words) during one minute. Word omissions, substitutions and mispronunciations were scored as errors, but self-corrections within 3 s after the error, repeated words, regionalisms, hesitations or words read slowly were not. Evidence of validity indicated significant differences in fluency between the four grades, and no differences were found between boys and girls. Test-retest reliability coefficients ranged between 0.91 and 0.97, and correlations with external criteria ranged between 0.24 and 0.94.

The data collection was authorized by the Portuguese Ministry of Education, school boards and parents. The assessments were conducted by the same psychologists that collect data in study 1. During the individual data collection that lasted between 10 and 20 min, students were assessed with the TLP and the TFL. The TCL was administered collectively in classrooms and lasted between 60 and 90 min.

4.1.3. Data analyses

To analyse the screening measure's classification accuracy, students have to be classified on an outcome measure as “passing” or “failing” (Clemens, Shapiro, & Thoemmes, 2011). In our study, two groups of at-risk readers were identified: a) classified as at-risk readers according to their scores on the reading fluency measure, and b) classified as at-risk readers based on the reading comprehension test results. Students who scored 1 standard deviation below the mean were classified as being at-risk readers. Other researchers have used a similar cut-off value (e.g., Catts, Fey, & Tomblin, 2001). Separate analyses were performed for oral reading fluency and for reading comprehension.

To determine the ability of each test form of the TLP to significantly predict the at-risk readers status, logistic regressions were computed. To examine the classification accuracy of the TLP-1, TLP-2, TLP-3 and TLP-4, receiver operating characteristic (ROC) curves' analyses were performed. As part of these analyses, two statistics were examined. First, the area under the ROC curve (AUC) was used to assess the overall index of classification accuracy of each test form. An AUC below 0.70 indicates low classification accuracy, between 0.70 and 0.80 is fair, 0.80 to 0.90 is good, and higher than 0.90 is excellent (Compton et al., 2006). Second, in order to define cut scores, we analysed the sensitivity and the specificity of each test form through the ROC curve. Following the recommendations of Jenkins et al. (2007) and Johnson and colleagues (2009), sensitivity levels were hold constant at 0.90 and 0.80 and then, the resulting specificity levels were analysed and the associated cut score selected. Note that in case of any cut score presenting a sensitivity level of approximately 0.90 or 0.80, we selected the immediate next cut score. In some cases, it corresponded to a sensitivity level of 1.00 when sensitivity was hold at 0.90. Next, using the cut score identified in the ROC analysis, we identified: the true positives (TP), the true negatives (TN), the false positives (FP), the false negatives (FN) and the hit rate (TP + TN / N).

Table 2
Reliability coefficients.

Test form	PSR	KR-20	ISR
TLP-1	0.91	0.92	0.99
TLP-2	0.88	0.92	0.99
TLP-3	0.82	0.86	0.98
TLP-4	0.74	0.82	0.97

Note. PSR = Person separation reliability; KR-20 = Kuder-Richardson formula 20; ISR = Item separation reliability.

4.2. Results

Descriptive statistics for the screening and the outcome measures are displayed in Table 3. The percentage of children who presented a performance below 1 SD on reading fluency was 8.1%, 18.7%, 15.3%, and 11.7% in grades 1, 2, 3, and 4, respectively. For the reading

Table 3
Descriptive statistics of the screening measure and the outcome measures.

	Overall			Risk determined by reading fluency performance						Risk determined by reading comprehension performance					
				At-risk			No risk			At-risk			No risk		
	n	M	SD	*n*	M	SD	*n*	M	SD	*n*	M	SD	*n*	M	SD
Screening															
TLP-1	82	105.44	11.4	7	83.0	10.4	75	107.5	8.9						
TLP-2	76	113.94	8.7	15	106.6	5.5	61	115.8	8.5	11	104.6	4.7	63	115.7	8.3
TLP-3	94	118.79	9.4	14	107.5	7.7	78	121.1	7.9	18	109.0	7.3	73	121.5	8.1
TLP-4	63	126.49	8.3	7	118.4	4.4	56	127.5	8.1	5	116.6	2.4	58	127.3	8.1
Outcome															
TFL G1	86	26.14	20.2	7	1.8	1.7	75	28.9	19.5						
TFL G2	79	65.72	25.3	15	32.5	7.4	64	73.5	21.3	11	39.9	16.6	65	70.9	23.8
TFL G3	92	92.03	35.1	14	41.1	11.2	78	101.2	29.6	18	56.3	22.7	73	100.7	32.2
TFL G4	65	113.48	28.6	8	66.9	8.0	57	120.0	23.4	5	93.0	6.0	58	116.3	29.0
TCL G2	76	14.28	4.8	14	10.5	3.9	62	15.1	4.6	11	7.2	1.4	65	15.5	4.0
TCL G3	91	16.38	5.4	14	11.1	4.7	77	17.3	4.9	18	8.2	2.3	73	18.4	3.7
TCL G4	63	19.30	4.5	7	18.4	1.6	56	19.4	1.6	5	9.2	2.8	58	20.2	3.4

Note. M = mean; SD = standard deviation; TLP = word reading measure; TFL = oral reading fluency measure; TCL = reading comprehension measure; G = grade.

comprehension measure the percentage was 12.2%, 19.7%, and 7.8% in grades 2, 3, and 4, respectively.

Results of the logistic regression analyses are displayed in Table 4. The results showed that each TLP test form was a significant predictor of each outcome variable.

Tables 5 and 6 display the results of the classification accuracy indices of the TLP test scores for each outcome variable. The AUC values were all above 0.70. Considering the guidelines by Compton et al. (2006), the data suggested that all the TLP test forms have good to excellent classification accuracy for at-risk readers considering either reading fluency or reading comprehension performances. Specifically, for reading fluency, the AUC values ranged between excellent in grade 1 and good in grades 2, 3 and 4, whereas for reading comprehension, the AUC values ranged between good in grades 2 and 3 and excellent in grade 4. When the sensitivity levels were set at 0.90, the recommended resulting specificity value of 0.80 was not reached when considering the reading fluency performance (see Table 5). The specificity rates were very low, ranging between 0.397 and 0.577 in grades 2 to 4. However, in grade 1, the specificity rate nearly reached the recommended value of 0.80. For the reading comprehension performance, the minimum value specificity of 0.80 was obtained in grades 2 and 4, but not in grade 3. Therefore, the percentage of accurate classification of students by the screen, considering the reading fluency performance, was very low, with percentages ranging between 49% and 63%, except in grade 1 that showed a satisfactory percentage of classification accuracy (i.e., 81%). On the contrary, the percentage of accurate classification of students by the screen, considering the reading comprehension performance, is satisfactory with a hit rate ranging between 85% and 86%, excepting grade 3, where a lower percentage of students was correctly classified.

When the sensitivity levels were hold at 0.80 for both outcome measures the resulting specificity rates should be at least at 0.70. This minimum specificity value was observed in all grades by the screen for both the outcome variables, except in grade 2 with a specificity value of nearly 0.70 for reading fluency (see Table 6). Consequently, the percentage of accurate classification of students by the screen, for both the outcome variables, was higher when the sensitivity level was set at 0.80. Regarding the reading comprehension performance, the hit rate in each grade was similar, ranging between 81% and 88%. Likewise, for the reading fluency measure, although the hit rates were lower, those values were also similar across grades, ranging between 72% and 78%, except in grade 1 with a very high classification accuracy (i.e., 94%).

In sum, considering the sensitivity, the specificity and the classification accuracy indices of the screening measure for both the outcome measures, setting the sensitivity levels at 0.80 produced better results.

5. Discussion

The aim of the first study was to develop and investigate the psychometric properties of the TLP. Then, in the second study, we investigated the classification accuracy of the TLP as a screening measure in classifying at-risk readers on two reading outcomes (reading fluency and reading comprehension). These analyses were performed for students from grades 1 to 4.

Regarding study 1, four forms of the TLP were developed for grades 1 to 4. Items that evidenced misfit and low point-measure correlations were excluded. The allocation of the remaining words in each test form was made according to the item difficulty level and to the students' word reading ability-level. Evidence of local independence of items and high to very high reliability coefficients was obtained for each test form.

Screening word recognition disabilities is relevant not only because poor word recognition skills are common in children with reading disabilities (Compton & Carlisle, 1994; Høien-Tengesdal & Tønnessen,

Table 4
Logistic regression analyses of the TLP test scores predicting outcome criterion measures.

Outcome criterion measure	Screening measure	*n*	B (S.E.)	Wald	*p*
TFL	TLP-1	82	−0.405 (0.132)	9.327	0.002
	TLP-2	76	−0.295 (0.085)	12.005	0.001
	TLP-3	92	−0.394 (0.093)	17.944	<0.001
	TLP-4	63	−0.505 (0.183)	7.643	0.006
TCL	TLP-2	74	−0.440 (0.127)	12.033	0.001
	TLP-3	91	−0.343 (0.079)	18.736	<0.001
	TLP-4	63	−0.793 (0.321)	6.109	0.013

Note. TFL = reading fluency measure; TCL = reading comprehension measure; S.E. = standard error.

Table 5
Classification accuracy indices for TLP test scores in identifying at-risk readers with sensitivity hold constant at 0.90.

Outcome criterion measure	Screening measure	AUC (S.E.)	95% IC	Sensitivity (≈0.90)	Specificity	Cut score	TP	TN	FP	FN	Classification accuracy
TFL	TLP-1	0.954 (0.030)	0.896–1.00	1.00	0.787	102.5	7	59	16	0	81%
	TLP-2	0.819 (0.056)	0.710–0.928	1.00	0.393	118	15	24	37	0	51%
	TLP-3	0.897 (0.050)	0.799–0.994	0.929	0.577	118	13	45	33	1	63%
	TLP-4	0.858 (0.067)	0.726–0.991	1.00	0.429	128	7	24	32	0	49%
TCL	TLP-2	0.898 (0.052)	0.797–1.00	0.909	0.841	108	10	53	10	1	85%
	TLP-3	0.875 (0.044)	0.791–0.961	0.944	0.603	118	17	44	29	1	67%
	TLP-4	0.928 (0.036)	0.856–0.999	1.00	0.845	120	5	49	9	0	86%

Note. TFL = reading fluency measure; TCL = reading comprehension measure; AUC = area under curve; S.E. = standard error; IC = interval confidence; TP = true positive; TN = true negative; FP = false positive; FN = false negative.

2011; Snowling & Hulme, 2005), but also because isolated word recognition is a predictor of success in reading acquisition (Verhoeven & van Leeuwe, 2009). Literature has suggested that screening measures should target reading skills that are congruent to the student's grade (Jenkins et al., 2007). Measures targeting word reading tasks, word reading plus reading fluency tasks, and reading fluency tasks, have been used as screens in first, second, and third grades, respectively (Johnson, Pool and Carter, n.d.). However, there are students that might still have difficulties in word reading in later grades (Lipka et al., 2006).

The TLP is a word reading screening measure that was developed to identify students who are at-risk for reading problems and who might benefit from reading intervention across elementary education. Our findings from study 2 suggest that the TLP is useful in identifying at-risk students for reading problems in grades 1 to 4. The results from ROC curve analyses indicated that the overall classification accuracy of each test form was good to excellent for students at-risk for reading fluency and comprehension problems. The TLP did not achieve the criterion for sensitivity and specificity of 0.90/0.80, when considering reading fluency performance as the outcome. The criterion was met only when considering reading comprehension performance as the outcome in grades 1, 2 and 4.

When using the criterion of 0.80/0.70 for sensitivity and specificity respectively, the screen achieved the criterion for the two outcome variables. The results showed that the TLP is well linked to the reading comprehension outcome, but not so well to the reading fluency variable, mainly when using the 0.90/0.80 criteria. These results are consistent with previous literature referring that "screens well linked to one criterion measure may not be well linked to another" (Jenkins et al., 2007, p. 584). Several studies (e.g., Adlof et al., 2006; Catts, Hogan, & Adlof, 2005) found that the influence of word recognition on reading comprehension decreases as schooling increases. Consequently, the possibility of using word recognition as a screen measure could be more appropriate in grades 1 and 2. In this second study, we demonstrated that it is possible to identify at-risk students in the four grade levels using a word recognition measure. However, this result should be further explored due to some limitations of this study. The number of at-risk readers is reduced in our sample, although their percentage in relation to the full sample is similar to the percentage of at-risk readers in other studies (Jenkins et al., 2007, for a review). Also the concurrent administration of the screening and the outcome measures might have overestimated the predictive classification accuracy (Jenkins et al., 2007). Other limitation is related with the absence of evidence that the TCL items are passage-independent, given that this feature has been reported as a matter of concern in some reading comprehension measures (e.g., Keenan & Betjemann, 2006).

6. Conclusion

The results concerning the development of the four forms of the TLP suggested that each test form presented satisfactory psychometric properties and might be used as a screening measure in the identification of students at-risk for reading problems. Given that Portuguese is the fifth language most spoken in the world, with approximatively 280 million speakers throughout the world, additional studies are needed to test if our results are replicable with Portuguese readers from other countries (e.g., Brazil, Angola).

Funding acknowledgements

This study was conducted at Psychology Research Centre (UID/PSI/01662/2013), University of Minho, and supported by the Portuguese Foundation for Science and Technology (FCT) and the Portuguese Ministry of Science, Technology and Higher Education through national funds and, when applicable, co-financed by the European Regional Development Fund (FEDER) through COMPETE2020 under the PT2020 Partnership Agreement (POCI-01-0145-FEDER-007653) and by Grant FCOMP-01-0124- FEDER-010733 from FCT and FEDER through the European program COMPETE (Operational Programme for Competitiveness Factors) under the National Strategic Reference Framework (QREN). The first author is also supported by grant from FCT (Grant SFRH/BD/78546/2011).

Table 6
Classification accuracy indices for TLP test scores in identifying at-risk readers with sensitivity hold constant at 0.80.

Outcome criterion measure	Screening measure	AUC (S.E.)	95% IC	Sensitivity (≈0.80)	Specificity	Cut score	TP	TN	FP	FN	Classification accuracy
TFL	TLP-1	0.954 (0.030)	0.896–1.00	0.857	0.947	92.5	6	71	4	1	94%
	TLP-2	0.819 (0.056)	0.710–0.928	0.867	0.689	111	13	42	19	2	72%
	TLP-3	0.897 (0.050)	0.799–0.994	0.857	0.769	116.5	12	60	18	2	78%
	TLP-4	0.858 (0.067)	0.726–0.991	0.857	0.768	122	6	43	13	1	78%
TCL	TLP-2	0.898 (0.052)	0.797–1.00	0.818	0.889	106.5	9	56	7	2	88%
	TLP-3	0.875 (0.044)	0.791–0.961	0.778[a]	0.822	115	14	60	13	4	81%
	TLP-4	0.928 (0.036)	0.856–0.999	0.800	0.879	118.5	4	51	7	1	87%

Note. TFL = reading fluency measure; TCL = reading comprehension measure; AUC = area under curve; S.E. = standard error; IC = interval confidence; TP = true positive; TN = true negative; FP = false positive; FN = false negative.

[a] Whenever that any cut-score was associated at a sensitivity level of 0.80, and when the immediate next cut-score above this value corresponded to the one used for the analyses with the sensitivity level of 0.90, we chose the cut-score immediately below the 0.80 value.

Appendix A. Distribution of the pool of words by psycholinguistic characteristic used in the development of the TLP.

Syllabic structure		High frequency				Low frequency				Total
		Regular		Irregular		Regular		Irregular		
		Short	Long	Short	Long	Short	Long	Short	Long	
1	CV·CV	5		2		4		2		13
1	CV·CV·CV		4		2		4		2	12
2	CVC·CV / CV·CVC	4		2		4		1		11
2	CVC·CV·CV / CV·CV·CVC		5		2		4		2	13
3	V·CV / CV·V	5		2		4		2		13
3	V·CV·CV CV·CV·V		5		2		4		2	13
4	CCV·CV / CV·CCV	4		2		4		1		11
4	CCV·CV·CV / CV·CV·CCV		4		2		4		1	11
5	CV·CVSW	5		2		4		2		13
5	CVG·CV·CV / CV·CV·CVSW		5		2		4		2	13
6	VC·CV / CV·VC	3		2		3		2		10
6	VC·CV·CV / CV·CV·VC		4		0		4		1	9
Total		26	27	12	10	23	24	10	10	142

Note. C = consonant; V = vowel; SW = semivowel.

References

Adlof, S. M., Catts, H. W., & Little, T. D. (2006). Should the simple view of reading include a fluency component? *Reading and Writing, 19*(9), 933–958. http://dx.doi.org/10.1007/s11145-006-9024-z.

Bacelar do Nascimento, M., Garcia Marques, M., & Segura da Cruz, M. L. (1987a). *Português fundamental, métodos e documentos, tomo 1, inquérito de frequência.* Lisboa: INIC, CLUL.

Bacelar do Nascimento, M., Rivenc, P., & Segura da Cruz, M. L. (1987b). *Português fundamental, métodos e documentos, tomo 2, inquérito de disponibilidade.* Lisboa: INIC, CLUL.

Berninger, V. W., Abbott, R. D., Vermeulen, K., & Fulton, C. M. (2006). Paths to reading comprehension in at-risk second-grade readers. *Journal of Learning Disabilities*, *39*(4), 334–351. http://dx.doi.org/10.1177/00222194060390040701.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications and programming.* New York: Routledge Academic.

Cadime, I., Ribeiro, I., Viana, F. L., Santos, S., Prieto, G., & Maia, J. (2013). Validity of a reading comprehension test for Portuguese students. *Psicothema*, *25*(3), 384–389. http://dx.doi.org/10.7334/psicothema2012.288.

Cadime, I., Ribeiro, I., Viana, F. L., Santos, S., & Prieto, G. (2014). Calibration of a reading comprehension test for Portuguese students. *Anales de Psicologia*, *30*(3), 1025–1034. http://dx.doi.org/10.6018/analesps.30.3.172611.

Castro, S. L., Caló, S., Gomes, I., Kay, J., Lesser, R., & Coltheart, M. (2007). *PALPA-P, Provas de Avaliação da Linguagem e da Afasia em Português [Tasks for the assessment of language processing and aphasia in Portuguese, PALPA-P].* Lisboa: CEGOC-TEA.

Catts, H. W., Fey, M. E., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools*, *32*(1), 38–50. http://dx.doi.org/10.1044/0161-1461(2001/004).

Catts, H., Hogan, T. P., & Adlof, S. M. (2005). Developmental changes in reading and reading disabilities. In H. Catts, & A. Kamhi (Eds.), *Connections between language and reading disabilities* (pp. 25–40) Mahwah, NJ: Erlbaum.

Clemens, N., Shapiro, E., & Thoemmes, F. (2011). Improving the efficacy of first grade reading screening: An investigation of word identification fluency with other early literacy indicators. *School Psychology Quarterly*, *26*(3), 231–244. http://dx.doi.org/10.1037/a0025173.

Compton, D. L., & Carlisle, J. F. (1994). Speed of word recognition as a distinguishing characteristic of reading disabilities. *Educational Psychology Review*, *6*(2), 115–140. http://dx.doi.org/10.1007/BF02208970.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*(2), 394–409. http://dx.doi.org/10.1037/0022-0663.98.2.394.

Daane, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., & Oranje, A. (2005). *Fourth-grade students reading aloud: NAEP 2002 special study of oral reading.* Washington, DC: Government Printing Office.

Francis, D. J., Fletcher, J. M., Catts, H., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris, & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 369–394).

Fuchs, L. S., Fuchs, D., & Compton, D. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children*, *71*(1), 7–21. http://dx.doi.org/10.1177/001440290407100101.

Goff, D. a., Pratt, C., & Ong, B. (2005). The relations between children's reading comprehension, working memory, language skills and components of reading decoding in a normal sample. *Reading and Writing*, *18*(7), 583–616. http://dx.doi.org/10.1007/s11145-004-7109-0.

Gomes, I., & Castro, S. L. (2003). Porlex: A lexical database in European Portuguese. *Psychologica*, *32*, 91–108.

Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi, & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 1–13). Mahwah, NJ: Lawrence Erlbaum.

Høien-Tengesdal, I., & Tønnessen, F. E. (2011). The relationship between phonological skills and word decoding. *Scandinavian Journal of Psychology*, *52*(1), 93–103. http://dx.doi.org/10.1111/j.1467-9450.2010.00856.x.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. http://dx.doi.org/10.1080/10705519909540118.

Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher*, *58*(8), 702–714. http://dx.doi.org/10.1598/RT.58.8.1.

Jenkins, J. R., & Johnson, E. (n.d). Universal screening for reading problems: Why and how should we do this? Retrieved July 14, 2016, from http://www.rtinetwork.org/essential/assessment/screening/readingproblems

Jenkins, J. R., & O'Connor, R. E. (2002). Early identification and intervention for children with reading/learning disabilities. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 99–149). New Jersey: Lawrence Erlbaum Associates, Publishers.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Accuracy and fluency in list and context reading of skilled and RD groups: Absolute and relative performance levels. *Learning Disabilities Research and Practice*, *18*(4), 237–245. http://dx.doi.org/10.1111/1540-5826.00078.

Jenkins, J. R., Hudson, R. F., & Johnson, E. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, *36*(4), 582–600.

Johnson, E., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research and Practice*, *24*(4), 174–185. http://dx.doi.org/10.1111/j.1540-5826.2009.00291.x.

Johnson, E., Pool, J., & Carter, D. (n.d). Screening for reading problems in grades 1 through 3: An overview of selected measures. Retrieved July 14, 2016, from http://www.rtinetwork.org/essential/assessment/screening/screening-for-reading-problems-in-grades-1-through-3.

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading*, *10*(4), 363–380. http://dx.doi.org/10.1207/s1532799xssr1004_2.

Klingbeil, D., McComas, J., Burns, M., & Helman, L. (2015). Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in the Schools*, *52*(5), 500–514. http://dx.doi.org/10.1002/pits.21839.

Lewandowski, L., Begeny, J., & Rogers, C. (2006). Word-recognition training: Computer versus tutor. *Reading & Writing Quarterly*, *22*(4), 395–410. http://dx.doi.org/10.1080/10573560500455786.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2011a). *A user's guide to WINSTEPS.* Chicago, IL: Winsteps.com.

Linacre, J. M. (2011b). *WINSTEPS.* Beaverton, OR: Winsteps.com.

Lipka, O., Lesaux, N. K., & Siegel, L. S. (2006). Retrospective analyses of the reading development of grade 4 students with reading disabilities: Risk status and profiles over 5 years. *Journal of Learning Disabilities*, *39*(4), 364–378. http://dx.doi.org/10.1177/00222194060390040901.

Lopes, F. T. F. (2011). *Dificuldades de escrita: O erro ortográfico, revelador do conhecimento metafonológico do escrevente aluno do ensino básico [Writing disabilities: Orthographic error and metaphonological knowledge of elementary graders].* Coimbra: Faculdade de Letras da Universidade de Coimbra.

Meisinger, E. B., Bloom, J. S., & Hynd, G. W. (2010). Reading fluency: Implications for the assessment of children with reading disabilities. *Annals of Dyslexia*, *60*(1), 1–17. http://dx.doi.org/10.1007/s11881-009-0031-z.

Muthén, B. O., & Muthén, L. (2010). *Mplus version 6.1 [Software].* Los Angeles, CA: Muthén&Muthén.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Rack, J. P., Snowling, M. J., & Olson, R. K. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly*, *27*(1), 29–53. http://dx.doi.org/10.2307/747832.

Ribeiro, I., Viana, F. L., Cadime, I., Chaves-Sousa, S., Santos, S., Silva, C., & Brandão, S. (2014). *Teste de Fluência da Leitura. Manual não publicado [The Test of Reading Fluency. Unpublished manual].* Braga: Escola de Psicologia: Universidade do Minho.

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

Slocum, T. (2002). Response to "Early identification and intervention for young children with reading/learning disabilities.". In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 179–184). New Jersey: Lawrence Erlbaum Associates, Publishers.

Snowling, M. J., & Hulme, C. (Eds.). (2005). *The science of reading: A handbook.* Oxford: Blackwell.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NY: Erlbaum.

Sucena, A., & Castro, S. L. (2012). *ALEPE - Avaliação da leitura em Português Europeu [ALEPE - Reading assessment in European Portuguese].* Lisboa: CEGOC.

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry*, *45*(1), 2–40. http://dx.doi.org/10.1046/j.0021-9630.2003.00305.x.

Verhoeven, L., & van Leeuwe, J. (2009). Modeling the growth of word-decoding skills: Evidence from Dutch. *Scientific Studies of Reading*, *13*(3), 205–223. http://dx.doi.org/10.1080/10888430902851356.

Viana, F. L., Ribeiro, I., Maia, J., & Santos, S. (2013). Propriedades psicométricas da Prova de Reconhecimento de Palavras [Psychometric properties of the Word Recognition Test]. *Psicologia: Reflexão e Crítica*, *26*(2), 231–240. http://dx.doi.org/10.1590/S0102-79722013000200003.

Wilkinson, G., & Robertson, G. (2006). *Wide range achievement test - Fourth Edition.* Lutz, FL: Psychological Assessment Ressources.

Woodcock, R. N. (2011). *Woodcock reading mastery tests - Third Edition.* San Antonio, TX: Pearson Assessments.