Psicologia: Teoria e Pesquisa
2020, v.36, e36325

DOI: https://dx.doi.org/10.1590/0102.3772e36325

Reliability, Validity and Standardization of the Reading Test: Sentence Comprehension*

Douglas de Araújo Vilhena[1, 2,**] & Ângela Maria Vieira Pinheiro[2]

[1]*Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil*

[2]*Universidade do Porto, Faculdade de Psicologia e de Ciências da Educação, Porto, Portugal*

ABSTRACT – The test called '*Reading Test: Sentence Comprehension* (TELCS)' has been validated and standardized. Participants (*N* = 1289, 2nd to 5th grade, 7 to 11-years-old) were stratified in 15 state-schools in Brazil. The TELCS demonstrated reliability and validity to classify reading performance by both school grade and chronological age. Correlations between the TELCS and a General Reading Composite score were high, as were those with reading accuracy rates of word and pseudoword. Cluster analysis suggested a five-class solution: reading disability, below, average, above, and high reading performance. For individual or collective use, TELCS can quickly screen the sentence reading ability, useful to identify those who might need additional support.

KEYWORDS: educational measurement, reading comprehension, reading disabilities, reading measures, evaluation

Fidedignidade, Validade e Normatização do Teste de Leitura: Compreensão de Sentenças

RESUMO – O teste chamado '*Teste de Leitura: Compreensão de Sentenças* (TELCS)' foi validado e normatizado. Participantes (*N* = 1289, 2º ao 5º ano, 7 a 11 anos) foram estratificados em quinze escolas estaduais. O TELCS demonstrou fidedignidade e validade para classificar o desempenho de leitura por ano escolar e idade cronológica. Houve alta correlação com o parâmetro composto de leitura geral, e as taxas de acurácia leitora de Palavras e de Pseudopalavras. Análise de *clusters* sugeriu cinco classes: transtorno, abaixo, média, acima, alta. De uso individual ou coletivo, o TELCS é um instrumento de rápido rastreio da habilidade de leitura de sentenças, útil para identificar aqueles que precisam de apoio adicional.

PALAVRAS-CHAVE: avaliação educacional, compreensão de leitura, distúrbios da leitura, medidas de leitura, avaliação

The Reading Test: Sentence Comprehension [*Teste de Leitura: Compreensão de Sentenças* (TELCS)] evaluates the reading comprehension ability. The TELCS is an adaptation of Lobrot's Lecture 3 reading test (L3) (Lobrot, 1967, 1980) to the Brazilian Portuguese language and socio-cultural context by Vilhena et al. (2016). The test demands efficient phonological and lexical word recognition, knowledge of individual word meaning, and executive functions, mainly working memory (Medina et al., 2018). The TELCS has been listed as a scientifically based tool to assess the reading comprehension capacity of children (Salles & Paula, 2016).

Background of the TELCS

The L3 test is used to evaluate the silent reading ability of French-speaking children and is part of the writing and reading ORLEC [*orthographe* (OR) and *lecture* (LEC)] battery proposed by Lobrot (1967, 1980). It consists of 36 incomplete sentences, followed by a choice of five words to complete each sentence. Only one of the five words is the correct answer (target word). The remainders are incorrect alternatives (distractors) and relate to the target word through visual, phonological, or semantic proximity or distance. The

* Apoio: Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais, CNPq (grant No. 134357/2013-2, and CAPES (Finance Code 001).

** E-mail: douglasvilhena@ufmg.br

■ Submetido: 05/09/2017; Revisado: 20/11/2019; Aceito:14/04/2020.

sentences are presented in an order of increasing difficulty (number of letters and syntactic complexity), with a time limit of five minutes to answer the test.

Since its creation, the ORLEC has been validated and updated norms are available (Génard et al., 1998; Mousty & Leybaert, 1999; Piérart & Grégoire, 2004). Mousty and Leybaert evaluated 217 monolingual French-speaking children in the 2nd and 4th school year in Belgium. The L3 test demonstrated good sensitivity for these grades, as no floor effect was observed in the second year (only 10% completed less than 5 items correctly) nor a ceiling effect in the fourth year (only 10% of children completed more than 30 items correctly). Piérart and Grégoire tested 2989 French-speaking Belgian elementary school children from the 3rd to 6th grade, provided updated norms for the L3, and demonstrated its high consistency and good reliability. As gender differences in scores in the 3rd and in the 5th grades were found, specific standardized and percentile norms for boys and girls were generated.

The L3 test has been used to evaluate reading ability in normal development (e.g. Rousselle & Noël, 2007) and in atypical conditions such as dyslexia (e.g. Serniclaes et al., 2004) and deafness (e.g. Alegría et al., 2009; Colin et al., 2013; Leybaert, 2000). The test has also been used as an exclusion criterion for children with reading difficulty (e.g. Faísca et al., 2019; Mussolin et al., 2010; Reybroeck & Hupet, 2009). The interest in L3, apart from its conceptual framework, psychometric properties, and ease of administration, is due to its existing adaptation for the Collective Test of Reading Efficacy [Test Colectivo de Eficacia Lectora (TECLE)] in Spanish (Carrillo & Marín, 2009) and the Reading Age Test [Teste de Idade de Leitura (TIL)] in European Portuguese (Sucena & Castro, 2010).

The TECLE (Carrillo & Marín, 2009; Marín & Carrillo, 1999) has been used since 1997 to screen for delayed reading in Castilian-speaking children. Although it has many similarities to the L3, it possesses a larger number of items (64 against 36), fewer alternative words for completing the sentence (four against five), and at least one pseudoword as a distractor for each item (in L3 all distractors are words). The TIL (Sucena & Castro, 2010) is more similar to the L3 than the TECLE, with the same number of items as the original French version and a similar structure. Normative data (score to Percentile) were provided to a sample of 614 children (8 to 11 years-old) and of 185 college students (18 to 48 years-old) (Fernandes et al., 2017; Sucena & Castro, 2010).

Adaptation of the L3 Test to Brazilian Portuguese

Reasons for departing mainly from the original French L3 test for use in Brazil, rather than from its European Portuguese adaptation, included the following issues: 1) the difference between the European and Brazilian Portuguese language as far as syntax and vocabulary are concerned; 2) lack of details in the available literature about the TIL in regard to the control of variables in its adaptation procedure; and 3) the fact that utilization of the original French version as the main reference would facilitate comparison between the three versions (L3, TIL, and TELCS).

The TELCS was adapted to Brazilian Portuguese from the original L3 test by Vilhena et al. (2016), who provided evidence of content validity by using the following steps. First, the sentences and the target words were translated from French to Brazilian Portuguese by two independent psychologists, proficient in both languages and knowledgeable about the test content. A conceptual rather than a strictly literal translation was done, taking into account the Brazilian cultural-linguistic context. Second, the distractors (incorrect alternatives) of the L3 Test were classified by their visual, phonological, or semantic proximity or distance to the target word, sentence, and other distractors; this classification was necessary because no detailed information was available in the published materials of the L3. Additionally, to prevent a given alternative guiding the response due to its greater familiarity, the selection of the Brazilian Portuguese distractors took into account the 'frequency of occurrence of words' using the Word Frequency Count in Written Brazilian Portuguese (Pinheiro, 2015).

For comparison purposes, these first two steps also took into consideration the TIL Test (the European Portuguese adaptation of the L3). The proximity between this version and the Brazilian was maintained as much as possible. In the third step, a blind reverse translation procedure, in which the translators – a Brazilian French teacher (also a psychologist) and a native French speaker highly proficient in Portuguese – had no access to the original version of the L3, confirmed the content equivalence.

In the fourth step, the level of proximity between the TELCS and L3 was ensured by maintaining in the Brazilian adaptation 26 sentences with the same meaning expressed in the L3's sentences, as well as the L3's length in number of letters (3598 letters in the L3 and 3284 in the TELCS). As for the proximity between TIL and L3, the Portuguese adaptation has 22 sentences with the same meaning as the L3 and it is shorter in length (3118 letters). The alteration of meaning in the remaining 14 items of TELCS in relation to the French original version varied from minor (n = 8) to major (n = 6) and in both cases the changes were due to ethical reasons (e.g. items with violent content) and to a search for precision or contextual adjustment (e.g. 'horse running' was changed to 'car race', as Brazilian children do not encounter horseracing).

The absence of Brazilian tests of reading comprehension poses problems to new test development in this area, because the lack of gold standards limits concurrent validation. Moreover, clinical practices face difficulty since the lack of validated tests induces intuitive and inadequate procedures that do not consider evidence-based practice that, among other purposes, recommend the integration of professional experience with scientifically proven knowledge to make the clinical exercise as objective as possible, granting efficacy and safety to evaluations and therapeutic interventions (El Dib, 2007).

Aside from the theoretical and practical relevance, establishing reliability, validity and standards for tools that measure reading comprehension of elementary school children is especially important in Brazil due to limitations of the currently available tests. To confirm that TELCS is an accurate measure of sentence reading comprehension, the present study was conducted for the following reasons: (a) to show evidence of reliability; (b) evidence of internal structure validity; (c) evidence of external validity; and (d) to provide standardized norms for 2nd to 5th grades and 7 to 11-years-old.

METHOD

Participants

All participants provided informed consent, and the Research Ethical Committee of the *Universidade Federal de Minas Gerais* approved all procedures in the study (identification number CAAE: 17754514.6.0000.5149), which was conducted in full accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki, 2008) for research involving human subjects.

Fifteen State Schools were stratified and randomly selected from a list provided by the Regional Superintendent of Education of all institutions registered in Belo Horizonte city, Brazil. Participants were children (*N* = 1289; 48% boys; age range = 7–11 years-old; 2nd to 5th grade), all native speakers of Brazilian Portuguese (see Table 1). The sample was collected at the end of the school year (November) of 2015 (*n* = 484, Table 2) and of 2018 (*n* = 805). A subgroup (*n* = 484; 49% boys; age range = 7–11; 2nd to 5th grade) (see Table 2), composed by six children (three boys and three girls) randomly selected from the attendance list from each of the 82 classrooms stratified across the city, completed a cognitive test battery of six instruments that evaluate reading ability, general cognitive ability, and social behaviour. The teachers (*N* = 82) completed a behaviour scale for each of the participants (*n* = 466). As there were no inclusion or exclusion criteria, an age range was found in each school grade due to natural birthday-determined variation and to grade retention (Table 1 and Table 2).

The required sample size was estimated taking into account the following parameters: a tolerance error of ±5%; Confidence Interval of 95.0%; population proportion of 0.5; and a target population size of 157.875 children enrolled in primary education in the city of Belo Horizonte. The suggested random sample size was 384 children. In addition,

Table 1

Sample distribution (N = 1289) by age groups, gender, and school grade

Age in years	Gender		School grade				Total
	Male	Female	2nd	3rd	4th	5th	
7	56	55	110	1	0	0	111
8	187	155	99	239	3	1	342
9	167	184	0	215	131	5	351
10	158	157	0	5	118	192	315
11	85	85	0	0	0	170	170
Total	653	636	209	460	252	368	1289

Table 2

Subgroup (N = 484) that completed a cognitive test battery, distribution by age groups, gender, and school grade

Age in years	Gender		School grade				Total
	Male	Female	2nd	3rd	4th	5th	
7	25	19	44	0	0	0	44
8	54	66	62	57	1	0	120
9	67	53	0	47	70	3	120
10	65	78	0	3	59	81	143
11	27	30	0	0	0	57	57
Total	238	246	106	107	130	141	484

the sample was increased to provide more power to the analyses, reaching 1289 participants, more than three times the required sample size.

Instruments

Six instruments were used in the external validity study – Reading Test: Sentence Comprehension (TELCS), Word Recognition Test (WRT), Pseudoword Recognition Test (PWRT), PROLEC's Reading Comprehension subtest (PROLEC-Text), Raven's Coloured Progressive Matrices Test (CPM), and the Strengths and Difficulties Questionnaire (SDQ).

The TELCS (Vilhena et al., 2016), printed on both sides of an A4 page, is composed of 36 isolated sentences (varying from 8 to 20 words), with the last word being always omitted. Each item is formed by five words that are displayed in a multiple-choice manner, with only one of them fitting the meaning to the sentence. The alternative words relate to each other in terms of visual similarities (e.g. number of letters, equal letters), phonological similarities (e.g. equal alliteration or rhyme), or semantic similarities (e.g. belonging to the same semantic category, such as type of profession). Other studies have provided to the TELCS additional evidence of content validity, internal structure validity (schooling effect), external validity (concurrent, criterion) (Machado & Maluf, 2019; Medina et al., 2018; Pinheiro, Vilhena & Santos, 2017; Vilhena & Pinheiro, 2016; Vilhena et al., 2016).

The Word Recognition Test (WRT) and the Pseudoword Recognition Test (PWRT) (Pinheiro, 2013) evaluate orthographic and phonological processing, respectively. Both WRT and PWRT consist of 4 training and 88 isolated test items each, which must be read aloud as quickly as possible. The words vary in frequency of occurrence, with equal numbers of high and low-frequency items. In each of these levels, the words further vary in grapheme–phoneme correspondence into regular and irregular words and in length (4 to 8 letters). The pseudowords were constructed with the same orthographic structure and length of stimuli used in the word test. Cogo-Moreira et al. (2012) demonstrated that WRT and PWRT correlated highly to each other ($r = .92, p < .001$) and with a text-reading accuracy measure ($r = .87–.92$; $p < .001$). Schooling effects, via Tobit regressions adjusted for the clusters of 10 schools, provided evidence of internal validation for WRT (third grade $\beta = 6.62$, $p < .01$; fourth grade $\beta = 10.56$, $p < .01$) and PWRT (third grade $\beta = 4.45$, $p < .001$; fourth grade $\beta = 6.77$, $p < .001$).

The Text Reading Comprehension subtest of the PROLEC Battery (PROLEC-Text) (Capellini et al., 2012) was used to evaluate semantic processes. It consists of four short texts, and of four literal questions about each one of them. Due to the type of questions the PROLEC-Text employs to assess comprehension, it will be considered here as a measure of literal comprehension. This construct is the first and most shallow level of text comprehension, with low engaging interactions with the text, as it requires extraction of information explicitly stated in a passage (Saadatnia et al., 2016).

Raven's Coloured Progressive Matrices Test (CPM) (Angelini et al., 1999) was used to measure general cognitive ability through the evaluation of analogical reasoning, which is the ability to infer relations between objects or elements (Pasquali et al., 2002). The test consists of 36 items, divided into three groups of 12 items (A, AB, B) organized with increasing difficulty. The task is to complete a figure at the top of a sheet with one of the six options printed below, which involves understanding that the images are characterized by their differences, similarities, identity, change, symmetry, and orientation.

Discriminant validity was assessed by the single-sided Brazilian version of the Strengths and Difficulties Questionnaire (SDQ) (Fleitlich, et al., 2000), a behavioural screening covering the age range 4 to 16, to be answered by the teachers. The instrument has 25 items divided into five scales: prosocial behaviour (empathy/positive relations), emotional symptoms (anxiety/mood), conduct problems (aggression/delinquency), hyperactivity/inattention, and peer relationship problems (withdrawn/social problems). The instrument has adequate indices of reliability and validity in 21 countries, including Brazil (Saur & Loureiro, 2012).

Procedures

Apart from the SDQ answered by the teachers during a period of one week, all instruments were administrated on the same day, in two sessions, , each lasting an average of 15 minutes. Whereas in the first session, groups of up to 10 children were assigned both the TELCS and CPM, in the second, each child was individually presented with both the WRT and PWRT (administrated in sequence, but in random order), followed by the PROLEC-Text. All instruments were administered by a professional psychologist and six undergraduate students of Psychology.

The TELCS was administered with a training phase composed of four items, with the first two answered collectively after being read aloud by the researcher and the other two items individually, via silent reading. The remaining 36 items were also read in silence by each child, however, as quickly as possible within a maximum of five minutes and with no assistance granted. The scoring of the test consisted of one point for each correct answer and zero for the incorrect or omitted ones (maximum: 36 points).

In both the WRT and PWRT, participants were asked to read aloud each item of each test card (printed on an A4 page, Arial font, size 14), starting from the first row from left to right. The reading time and errors were registered. Time to read WRT ranged from 48 seconds to six minutes (average of 127 seconds) and PWRT from 53 seconds to nine minutes (average of 175 seconds). On both instruments, two measures were used: accuracy, which is the total number of correctly read words or pseudowords, and accuracy rate, which is the total number of correct words or pseudowords read per minute.

Regarding the PROLEC-Text, the stories were administrated in a fixed order, after the following statement: 'I will display a small text for you to read. Read it carefully because after you finish I will ask you some questions about it.' The participant was asked to read each story quietly, without time limit, and to respond orally to open questions (also made orally), immediately after reading each text. No rereading was allowed.

The CPM was individually administrated to 2nd year participants and the collective form was used for children from grades 3 to 5. It was presented as a puzzle game: the first two items were introduced collectively and explicitly, with subsequent items answered without assistance. There was no time limit. No child spent more than 12 minutes to complete the test.

Data Analysis

All analyses were performed using SPSS version 21.0 (IBM, Chicago, Illinois). No outliers were detected using the outlier labeling rule with a *g* value of 2.2. Test reliability was analysed by Cronbach's alpha, Spearman-Brown split-half coefficient and Test-Retest (internal consistency indexes). Internal validity (schooling and age effect) was assessed by hierarchical two-step cluster analysis and by univariate analysis (ANOVA), corrected for family-wise error with Bonferroni. For the investigation of the TELCS' internal validity and population distribution scores (Figures 1 and Figure 2), skewness and kurtosis values were divided by the respective standard error, using a criterion of significance higher than 1.96 (Cramer & Howitt, 2004).

External validity (relations between the TELCS and all reading measures, general cognitive ability, behaviour, and demographic variables) was assessed by Pearson Correlations. Since the TELCS evaluates reading competence as a whole, a dimension reduction by principal components analysis (Carreira-Perpiñán, 1997) was performed to incorporate three reading measures (PROLEC-Text and accuracy rates of the WRT and PWRT) to create a robust reading variable, the General Reading Composite.

In order to classify a given participant according to his or her reading ability, several statistically distinct latent groups were identified in the sample via hierarchical two-step cluster analysis. This method assumes that the distance between two clusters is equivalent to the decrease in log-likelihood function as a result of merging. The Bayesian information criterion (BIC) was established to compare the number of latent classes, in which small values correspond to a better fit. Statistical significance was set at $p < .05$.

RESULTS

Evidence of Reliability

The strong internal consistency shown by Cronbach's alpha index (.95) and Spearman–Brown split-half coefficient (.97) provided evidence of the reliability of the TELC. Test-Retest reliability demonstrated that TELCS's mean scores were stable between conditions, as there was no difference between the samples collected at 2015 and 2018, according to both school grades [2nd ($F_{(1, 207)} = .1, p = .97$), 3rd ($F_{(1, 459)} = 2.1, p = .15$), 4th ($F_{(1, 250)} = 1.6, p = .21$), and 5th grade ($F_{(1, 365)} = 2.7, p = .10$)]; and to chronological age [7 ($F_{(1, 109)} = .5, p = .46$), 8 ($F_{(1, 340)} = 2.0, p = .15$), 9 ($F_{(1, 349)} = 1.8, p = .29$), 10 ($F_{(1, 313)} = 2.2, p = .14$), and 11 years-old ($F_{(1, 168)} = .7, p = .41$)]. Thus, the samples from 2015 and 2018 could be merged for the standardization study.

Evidence of Internal Structure Validity

Evidence of the internal structure validity of the TELCS was demonstrated by significant schooling [$F_{(3, 1285)} = 200.0$, MSE = 57.5, $p < .001$, (2nd < 3rd < 4th < 5th grade)], and age effects [$F_{(4, 1284)} = 136.9$, MSE = 59.2, $p < .001$, (7 < 8 < 9 < 10 < 11 years-old)]. TELCS's scores with 95% confidence interval did not overlap between school grades (2nd grade: 8.1–9.8, 3rd grade: 13.5–14.9, 4th grade: 19.1–21.1, 5th grade: 22.6–24.3) nor between chronological ages (7 years: 6.8–9.0, 8 years: 12.3–13.9, 9 years: 15.2–16.9, 10 years: 21.0–22.7, 11 years: 23.6–26.0).

The sample (N = 1289) showed a standard normal distribution (Figure 1.a), with symmetric skewness (.06). In the split condition by school grade and age, significant skewness was found in for the 2nd (0.60) and the 5th (-0.55) grades, and for 7 (0.45), 8 (0.37), 10 (-0.34), and 11 (-0.66) years-old. The curve demonstrated a platykurtic curve (-0.79) (Figure 1.a), with significant kurtosis found for the 4th grade (-0.68), and for 9 years-old (-0.53). Using the floor effect criterion of less than or equal to five points in the TELCS (used by Mousty and Leybaert, 1999), poor performers were found in the 2nd (31.1%), 3rd (13.9%), 4th (3.6%), and 5th grade (3.0%), and in participants aged 7 (38.7%), 8 (16.4%), 9 (10.5%), and 10 years-old (2.9%), but not in those aged 11 years-old (all 170 participants scored more than 5 points).

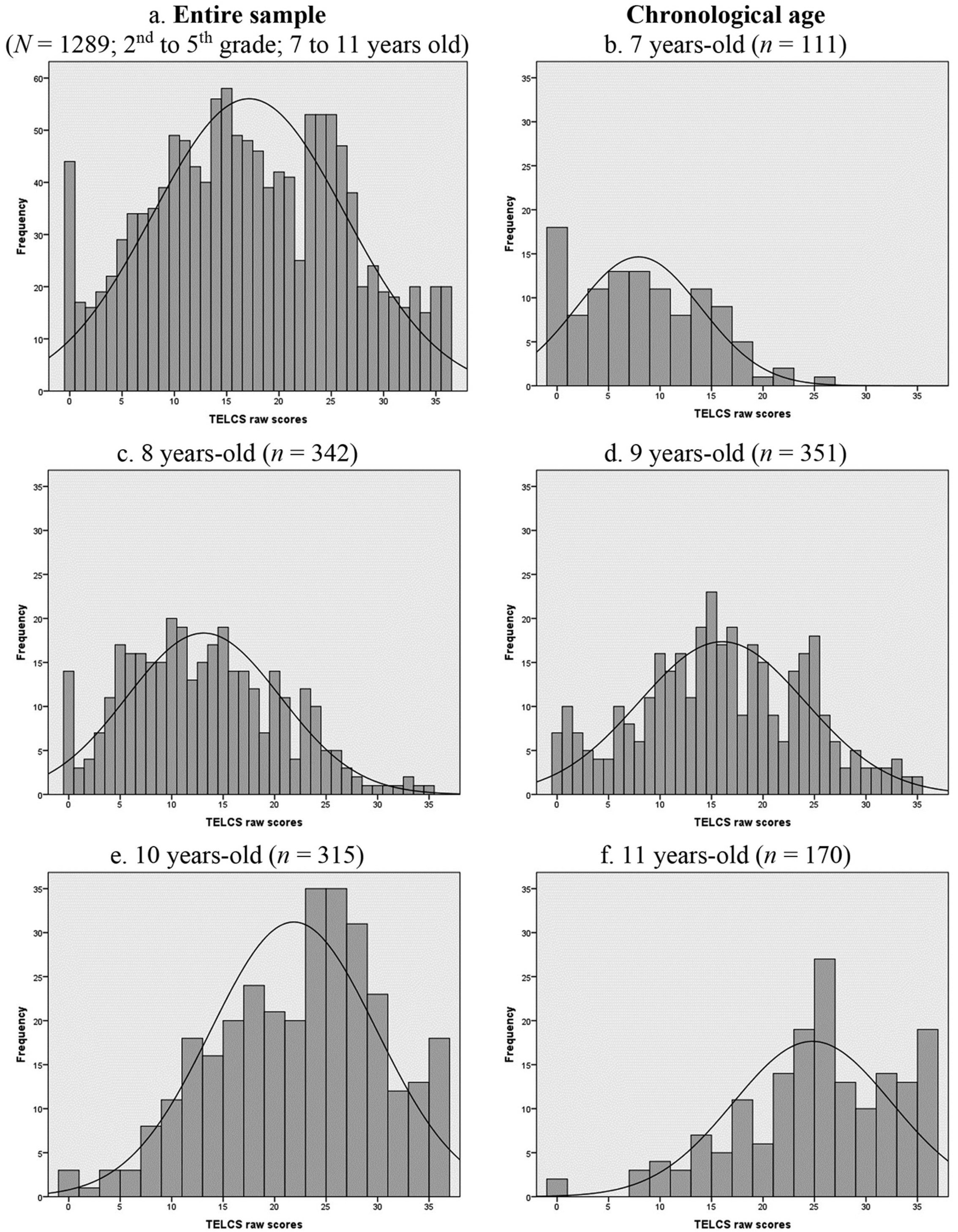

Figure 1. TELCS scores distribution for the entire sample (Figure 1.a) and for each age group (Figure 1.b–f), fitted with an expected normal distribution curve.

Evidence of External Validity

For concurrent external validity of the TELCS, a bivariate correlation showed a significant *p*-value of .001 for all reading and general cognitive ability tests (Table 3). There were strong correlations with accuracy rates of Word and Pseudoword Recognition Test (*r* = .84 and .79, respectively) and with the General Reading Composite (*r* = .84); moderate correlations with the untimed accuracy measures of the Word and Pseudoword Recognition Test (*r* = .55 and .57), with the PROLEC-Text (*r* = .58), and with general cognitive ability (*r* = .50). In contrast, the external discriminant validity of

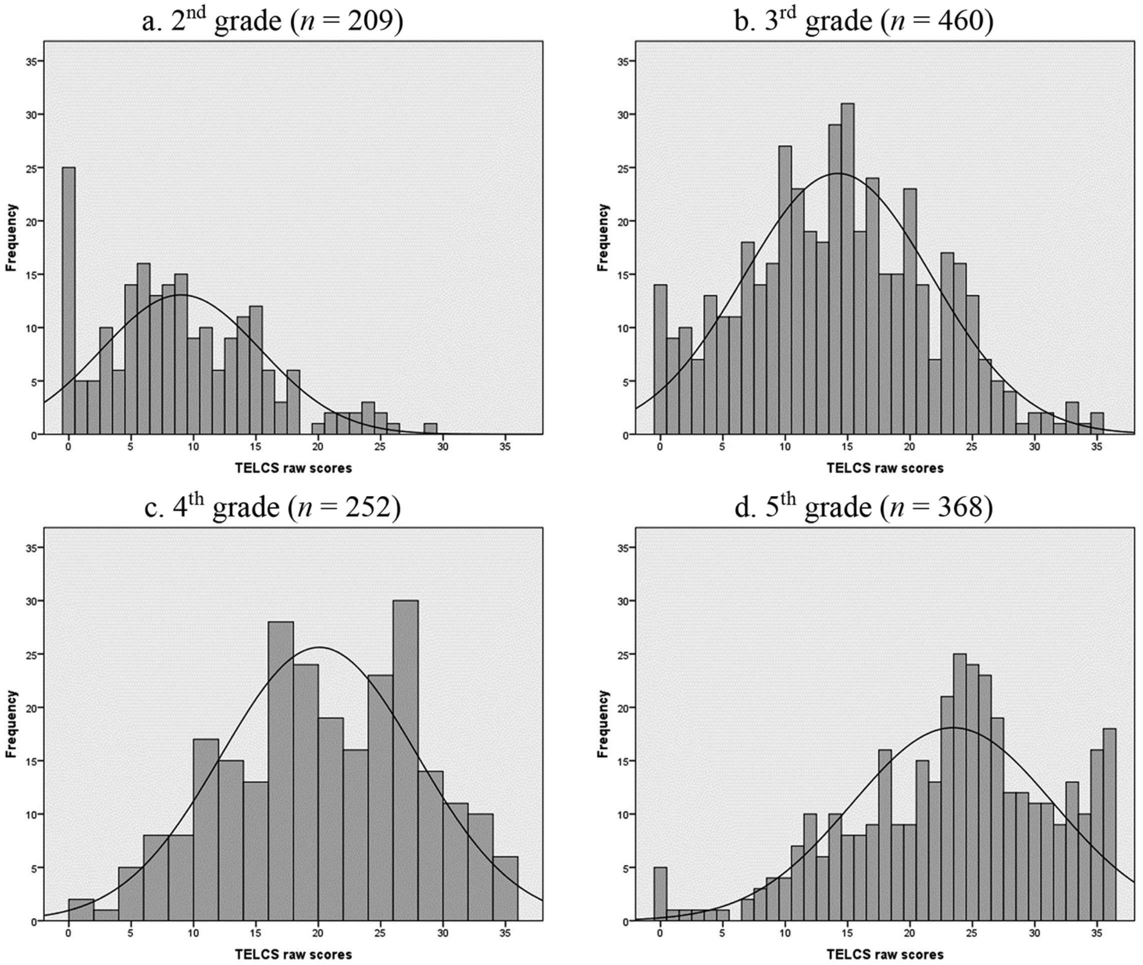

Figure 2. TELCS scores distribution for each school grade (Figure 1.a–d), fitted with an expected normal distribution curve.

Table 3

Pearson Correlations between Reading, General Cognitive Ability, Behaviour, and Demographic Variables

			TELCS	Grade	Age
Reading	TELCS		.566**	.565**	
	WRT	Accuracy	.550**	.330**	.319**
		Accuracy rate	.840**	.557**	.570**
	PWRT	Accuracy	.570**	.306**	.297**
		Accuracy rate	.787**	.509**	.526**
	PROLEC-Text		.582**	.384**	.380**
	General reading composite		.837*	.528*	.546*
Cognition	CPM		.502**	.445**	.432**
Behaviour (SDQ)	Prosocial Behaviour		.161*		
	Emotional Symptoms		-.290*		
	Conduct Problems		-.216*		
	Hyperactivity/Inattention		-.375*		
	Peer Relationship Problems		-.126*		
	Total negative behaviours		-.344*		

Note. $^{*}p < .01$ (2-tailed), $^{**}p < .001$ (2-tailed). SDQ: Strengths and Difficulties Questionnaire; TELCS: Reading Test: Sentence Comprehension; WRT: Word Recognition Test; PWRT: Pseudoword Recognition Test; PROLEC-Text: PROLEC Text Comprehension subtest; CPM: Raven's Coloured Progressive Matrices Test. Pearson correlations between SDQ, Grade, and Age were below .10 and were omitted from the Table.

the TELCS was provided by a mild correlation found with psychiatric behaviours (r = -.34). As seen in Table 3, the TELCS was the variable with the highest correlation with school grade and age, even in comparison with CMP, which has a strong age correspondence due to neural maturation. The analysis of variance showed no difference between genders for the TELCS ($F_{(1, 1288)} = 1.32$, $p = .25$).

Standardization Study

Table 4 presents the norms for the TELCS' standardization study (N = 1289 participants), with raw scores, corresponding percentiles, reading performance classification, and descriptive statistics. Distinct groups delineated by reading performance in the standardization were supported by a two-step cluster analysis, which suggested a five-class solution: reading disability; low; average; above average; and high performance. An univariate analysis of variance (ANOVA), with a Bonferroni correction, confirmed significant differences in scores for all five classes ($p < .001$).

Table 4

Standardization Study – TELCS's raw score to percentile and to reading performance classification, with descriptive statistics according to the participant's school grade and chronological age

Reading performance	**Percentile**	**School grade**				**Chronological age in years**				
		2nd	***3rd***	***4th***	***5th***	***7***	***8***	***9***	***10***	***11***
Disability	7	0	2	8	11	0	3	3	9	12
	10	0	3	10	12	0	4	5	11	14
Low	15	2	6	11	14	0	5	7	12	17
	25	5	9	15	18	4	7	10	16	21
Average	30	6	10	16	20	5	8	12	18	22
	40	7	12	18	23	6	10	14	20	24
	50	8	14	20	24	7	13	16	23	25
Above average	60	10	16	23	26	9	15	18	25	27
	70	12	18	25	28	11	17	20	27	29
	80	14	21	27	31	14	20	24	29	32
High	90	17	24	30	34	16	23	26	32	35
	95	22	26	32	35	18	26	29	35	36
Mean	-	9.0	14.2	20.1	23.4	7.9	13.1	16.0	21.9	24.8
Standard Deviation	-	6.4	7.5	7.8	8.1	6.0	7.4	8.1	8.1	7.7

DISCUSSION

The purpose of this study was to provide evidence of reliability, internal and external validity, as well as standardized norms for the *Reading Test: Sentence Comprehension* [Teste de Leitura: Compreensão de Sentenças]. Evidence of content validity was provided by Vilhena *et al*. (2016), with detailed description of the operational and constitutive definitions of the items.

Evidence of Reliability

The TELCS presented strong reliability indices ($\alpha = .95$; $\rho = .97$), which were very close to those found for the original L3 ($\alpha = .94$; $\rho = .98$) (Piérart & Grégoire, 2004). Another evidence of reliability was provided by Test-Retest measure taken in 2015 and 2018, that demonstrated the stability of TELCS' mean scores between conditions, partially due to the representative stratified random sampling.

Evidence of Internal Validity

Regarding validity evidence-based on the internal structure, the instrument significantly distinguishes readers both by school grade (2nd < 3rd < 4th < 5th grade) and by chronological age (7 < 8 < 9 < 10 < 11 years-old). The differences were considered significant, as the data presented non-overlapping confidence intervals. Machado and Maluf (2019) also confirmed TELCS's schooling effect ($F_{(3,95)} = 41.2$, $p < .001$) in their sample (N = 98): 2nd [Mean = 9.6, SD = 6.3] < 3rd [M = 18.2, SD = 5.8] < 4th grade [M = 23.2, SD = 6.7].

The results demonstrated the standard normal distribution of the data. The significant levels of skewness shown in the 2nd and 5th grades were due to floor and ceiling effects, respectively. An alarming result was the identification of children who performed poorly within the sample. It was found that 31% of the Brazilian 2nd graders completed less than 5 items correctly, a much worse performance than the 10% reported by Mousty and Leybaert (1999) for the same level in Belgium. On the other hand, 11.5% of the current sample of 4th graders completed more than 30 items correctly, similar to the 10% found by Mousty and Leybaert in the fourth school year. However, 24.0% in the 5th grade completed more than 30 items correctly, evidence of ceiling effect. In the future, a reduction of the TELCS's examination time from five to four minutes (or less) should be tested to control its present performance overestimation for 5th graders.

Evidence of External Validity

Evidence based on relationships with external variables includes concurrent, discriminant and criterion validity. For concurrent external validity, as expected, the TELCS presented moderate to strong correlations with all the comparison reading instruments. The strongest correlation (*r* = .84) was with the accuracy rate for word recognition and the General Reading Composite. Vilhena and Pinheiro (2016) found a moderate correlation (*r* = .65) between TELCS and the Scale of Evaluation of Reading Competence by the Teacher (EACOL), an indirect assessment of reading aloud (speed and accuracy in word recognition, prosody, and comprehension) and silent reading (text comprehension and synthesis) of schoolchildren. D'Hondt and Leybaert (2003) also reported a significant correlation with the L3 using a timed lexical decision task (*r* = .65, *p* < .001). These results agree with the verbal efficiency theory, which states that poor word representations and slow decoding processes (mapping orthographic to phonological representations) consume resources in working memory that would otherwise be dedicated to high-level comprehension processes (Hamilton et al., 2016; Perfetti, 1985). Quick word recognition is especially important in the TELCS, as working memory is already being engaged to complete sentences by selecting a target word among five options.

As Brazil has no gold standard sentence reading comprehension test that could be used to establish the concurrent external validity of the TELCS, it was necessary to create, via a dimension reduction technique, a General Reading Composite score to integrate the three reading measures employed in the present study (WRT accuracy rate; PWRT accuracy rate; and PROLEC-Text) into a single and comprehensive variable. The high correlation between TELCS and General Reading Composite (*r* = .84) can be considered the most important result of the current study, providing evidence of concurrent external validity.

Nonverbal intelligence has consistently been demonstrated to be a mild–moderate predictor of reading comprehension in children (Kershaw & Schatschneider, 2012; Stanovich *et al.*, 1984). High correlations between the CPM and reading are not expected because this would tend to eliminate the causal influence of reading upon reading ability (Carver, 1990). In light of the referred verbal efficiency theory (Perfetti, 1985), intelligence mainly influences tests that evaluate reading efficiency, a combination of accuracy and rate measures. Carver found moderate correlations between the CPM and a reading efficiency test in 2–12 graders (range = .36 to .68, mean = .49). These values agree with the correlations found in the current study, as general cognitive ability played a moderate role in TELCS scores (*r* = .50).

Further evidence of concurrent external validity was provided by Medina et al. (2018) that demonstrated, in a sample of 20 children with and without diagnosis of dyslexia, that the TELCS had a strong correlation with the reading measure (Test for School Achievement, *r* = .93), strong correlation with phonological awareness (Phonemic Awareness Tasks, *r* = .79), and moderate to strong correlation with different components of executive functions, such as Cognitive Flexibility (Test of Tracks A: *r* = .51; Tracks B: *r* = .74), Working Memory (Working Memory Tasks: *r* = .82; Visuospatial Task: *r* = .58; Digits: *r* = .75; Phonological Span of Nonwords: *r* = .74), Inhibitory Control (Attention by Cancelling Test: *r* = .49; Task Go/No Go: *r* = .55 to .58), and with Verbal Fluency (*r* = .59).

Concerning the psychiatric aspects, signs of behavioural problems, indexed by the SDQ, showed a negative effect on TELCS scores. This agrees with Kristoffersen *et al.* (2014) who reasoned that a negative association between indicators of externalizing behaviour and school outcomes can be expected. In contrast, prosocial behaviour, also measured by the SDQ, was beneficial to the reading performance of the participants. The weak correlations between reading and these negative/positive social behaviours provide evidence of discriminant validity for the TELCS, as they measure different framework constructs.

Criterion validity would be provided if TELCS could identify children with dyslexia. One possibility is to compare a group of children with such diagnosis to a group of children without complaints of reading and writing difficulties. Medina *et al.* (2018) verified that the group of dyslexic children (Mean = 1.4, SD = 0.83) had a significant poorer performance in TELCS, when compared to the control group without reading difficulty (Mean = 21.8; SD = 2.15) (*Mann-Whitney* = 0.0, *p* < 0.001). Likewise, Medina and Guimarães (2019) demonstrated that the two groups of dyslexic children had the same score in TELCS (*p* = .44), and that both dyslexic groups had poorer scores than two control groups composed by good readers matched for age and younger (*p* > .05). Medina and Guimarães also verified that the TELCS was able to detect the development of reading (higher means) throughout the school year ($p < .05$) for the dyslexic children, probably due to the advancement in decoding transferred to the sentence comprehension. These results offer evidence that TELCS has criterion validity, being able to discriminate the diagnostic group, as dyslexic groups showed poorer performance in TELCS, with lower scores.

Standardization study

For the standardization study, the TELCS' norms were split into school grades and chronological age, as years of study reflect formal schooling stimulation, and age is an indicator of neural maturation. The large random stratified sample size ($N = 1289$) can be regarded as one of the strengths of the study. However, it is important to highlight that it is a regional sample and may not represent the performance of Portuguese-speaking readers across the country. No difference between genders was found in the current study ($p = .25$). Using the TELCS, Machado and Maluf (2019) also did not find gender difference in the 2nd ($p = .48$), 3rd ($p = .92$), and 4th grades ($p = .75$). Thus, the norms of the present study were not split into male and female as was done in both Piérart and Grégoire (2004) and Sucena and Castro's (2010) standardization studies. A reanalysis of Sucena and Castro's data suggested that the gender difference reported was, in fact, concentrated only in the 4th grade ($F_{(1,\ 123)} = 2.78$, $p = .006$, Cohen's $d = .50$) and not in all grades.

The large number of items in the test ($N = 36$) permitted a clear differentiation and classification of participants' reading performance (reading disability, low, average, above average, and high performance), with a standard normal distribution curve. Few participants in the entire sample (1.6%) could finish the test in five minutes with 100% accuracy, which may be an indication of its adequate length. In summary, from the results obtained, it can be asserted that the TELCS is reliable for the assessment of different levels of reading ability in children from the 2nd to the 4th grade. The ceiling effect found in the 5th grade reveals that the TELCS has limitations in discriminating the reading performance of children at advanced levels of schooling.

TELCS grade and age scores were divided according to different parameters of performance, allowing researchers and clinicians to select a lenient or conservative cut-off score according to their purposes (cut-offs at 25th, 15th, 10th, and 7th Percentiles). This range of parameters was based on literature. Génard *et al*. (1998), demonstrated that 69 out of 75 dyslexic children scored in the lowest quartile on the L3, and thus considered the 25th percentile to be a good predictor of reading disability, especially for research purposes (higher sensitivity). Rousselle and Noël (2007) asserted that the choice of the 15th percentile not only guarantees the diagnosis of reading disability, but also avoids false positives when used for clinical purposes (higher specificity). Taking a more rigorous step, the Diagnostic and Statistical Manual-5 (DSM-5) (American Psychiatric Association, 2013) recommends a cut-off score in the 7th percentile for a Specific Learning Disorder (in this case, with the specifier for impairment in reading). However, when considering an academic skill much below average age, this manual also endorses a more lenient threshold of up to the 25th percentile.

TELCS meets the standards for reliability and validity. The current standardization study involved a large, representative, stratified, and random sample ($N = 1289$). The results offered here are limited to Portuguese-speaking children attending local public schools, taking into consideration the city of Belo Horizonte. One of the implications of the gloomy picture portrayed by the figures reported in the present study is the need to develop norms not only for those children attending both public and private schools in other regions of Brazil, but also for college students, followed by the strengthening of the reliability and validity of the test to evaluate the sentence reading comprehension ability.

CONCLUSION

The TELCS has demonstrated adequate psychometric properties, with evidence of reliability, internal structure validity, content validity, external validity (concurrent, discriminant, criterion) and standardized scores for measuring sentence reading comprehension for Brazilian Elementary School children (2nd through 5th grades; 7 to 11 years-old). Due to the psycholinguistic controls introduced by Vilhena *et al*. (2016), TELCS can be used to screen children with low to high reading performance, either for collective screening purposes or for individual clinical administrations. These features make the TELCS an important psychometrically standardized measure to assess the Criterion B for Specific Learning Disorder (specifier for impairment in reading, also referred to as Dyslexia) in the DSM-5, which requires an academic skill substantially and quantifiably below those expected for the individual's chronological age.

REFERENCES

Alegría, J., Domínguez, A. B., & Straten, P. (2009). ¿Cómo leen los sordos adultos? La estrategia de palabras clave [How do deaf adults read? The key word strategy]. *Revista de Logopedia, Foniatría y Audiología, 29*(3), 195–206. https://doi.org/10.1016/S0214-4603(09)70028-2

Angelini, A. L., Alves, I. C. B., Custódio, E. M., Duarte, W. F., & Duarte, J. L. M. (1999). *Matrizes Progressivas Coloridas de Raven: Escala Especial.* CETEPP.

APA (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.). American Psychiatric Association.

Capellini, S. A., Oliveira, A. M., & Cuetos, F. (2012). *PROLEC: Provas de Avaliação dos Processos de Leitura* (2nd ed.). Casa do Psicólogo.

Carreira-Perpiñán, M. Á. (1997). *A Review of Dimension Reduction Techniques.* Technical Report. Department of Computer Science*,* University of Sheffield.

Carrillo, M. S., & Marín, J. (2009). Test de Eficiencia Lectora – TECLE. In A. Cuadro, D. Costa, D. Trias, & P. Ponce de León (Eds.), *Evaluación del nivel lector: Manual técnico del test de eficacia lectora (TECLE)*. Prensa Médica Latinoamericana.

Carver, R. P. (1990). Intelligence and reading ability in grades 2-12. *Intelligence*, *14*(4), 449–455. https://doi.org/10.1016/S0160-2896(05)80014-5

Cogo-Moreira, H., Ploubidis, G., De Avila, C., Mari, J., & Pinheiro, A. M. V. (2012). EACOL (Scale of Evaluation of Reading Competency by the Teacher): Evidence of concurrent and discriminant validity. *Neuropsychiatric Diseases and Treatment*, *8,* 443–454. https://doi.org/10.2147/NDT.S36196

Colin, S., Leybaert, J., Ecalle, J., & Magnan, A. (2013). The development of word recognition, sentence comprehension, word spelling, and vocabulary in children with deafness: A longitudinal study. *Research in Developmental Disabilities*, *34*, 1781–1793. https://doi.org/10.1016/j.ridd.2013.02.001

Cramer, D., & Howitt, D. (2004). *The Sage Dictionary of Statistics: A practical resource for students in the social sciences.* Thousand Oaks: Sage.

Declaration of Helsinki (World Medical Association). (2008). 7th Revision of the Declaration of Helsinki: Ethical principles for medical research involving human subjects. 59th World Medical Association (WMA) General Assembly, Seoul, South Korea.

D'Hondt, M., & Leybaert, J. (2003). Lateralization effects during semantic and rhyme judgement tasks in deaf and hearing subjects. *Brain and Language*, *87*(2), 227–240. https://doi.org/10.1016/S0093-934X(03)00104-4

El Dib, R. P. (2007). Como praticar a medicina baseada em evidências [How to practice evidence-based medicine]. *Jornal Vascular Brasileiro*, *6*(1), 1-4. http://dx.doi.org/10.1590/S1677-54492007000100001

Faísca, L., Reis, A., & Araújo, S. (2019). Early Brain Sensitivity to Word Frequency and Lexicality During Reading Aloud and Implicit Reading. *Frontiers in psychology*, *10*, 830. https://doi.org/10.3389/fpsyg.2019.00830

Fernandes, T., Araujo, S., Sucena, A., Reis, A., & Castro, S. L. (2017). The 1-min Screening Test for Reading Problems in College Students: Psychometric Properties of the 1-min TIL. *Dyslexia*, *23*(1), 66-87. https://doi.org/10.1002/dys.1548

Fleitlich-Bilyk, B., Cortázar, P. G., & Goodman, R. (2000). Questionário de Capacidades e Dificuldades (SDQ) [Strengths and Difficulties Questionnaire (SDQ)]. *Infanto: Revista de Neuropsiquiatria da Infância e Adolescência*, 8(1), 44-50.

Génard, N., Mousty, P., Content, A., Alegria, J., Leybaert, J., & Morais, J. (1998). Methods to establish subtypes of developmental dyslexia. In P. Reitsma & L. Verhoeven (Eds.), *Problems and interventions in literacy development* (pp. 163–176). Neuropsychology and Cognition, vol 15. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-2772-3_10

Hamilton, S., Freed, E., & Long, D. L. (2016). Word-Decoding Skill Interacts with Working Memory Capacity to Influence Inference Generation During Reading. *Reading Research Quarterly*, *51*(4), 391-402. https://doi.org/10.1002/rrq.148

Kershaw, S., & Schatschneider, C. (2012). A latent variable approach to the simple view of reading. *Reading and Writing*, *25*(2), 433-464. https://doi.org/10.1007/s11145-010-9278-3

Kristoffersen, J. H. G., Obel, C., & Smith, N. (2014). Gender differences in behavioral problems and school outcomes. *Journal of Economic Behavior & Organization, 115*, 75-93. https://doi.org/10.1016/j.jebo.2014.10.006

Leybaert, J. (2000). Phonology acquired through the eyes and spelling in deaf children. *Journal of Experimental Child Psychology*, *75*(4), 291–318. https://doi.org/10.1006/jecp.1999.2539

Lobrot, M. (1967). *Batterie pour mesurer la lecture et l'orthographe, ORLEC*. Beaumont/Oise: Bureau d'études et de recherches.

Lobrot, M. (1980). *Lire avec e 'preuves pour e 'valuer la capacite 'de lecture (D-OR-LEC)*. Editions ESF.

Machado, M. S. M., & Maluf, M. R. (2019). Como evolui a compreensão da leitura em alunos do ensino fundamental [How reading comprehension evolves in elementary school students]. *Psicologia da Educação*, (49), 57-66. https://doi.org/10.5935/2175-3520.20190019

Marín, J., & Carrillo, M. S. (1999). *Test Colectivo de Eficacia Lectora (TECLE).* Unpublished Manuscript. Departamento de Psicología Básica y Metodología. Universidad de Murcia. Available from https://www.um.es/langpsy/Publicaciones/TECLE_Marin%20y%20Carrillo%20(1999).pdf

Medina, G. B. K., & Guimarães, S. R. K. (2019). Reading of Students with Developmental Dyslexia: Impacts of an Intervention with Phonic Method Associated with the Executive Functions Stimuli. *Revista Brasileira de Educação Especial*, *25*(1), 155–174. http://dx.doi.org/10.1590/s1413-65382519000100010

Medina, G. B. K., Souza, F. F., & Guimarães, S. R. K. (2018). Funções executivas e leitura em crianças brasileiras com dislexia do desenvolvimento [Executive functions and reading in Brazilian children with developmental dyslexia]. *Revista Psicopedagogia*, *35*(107), 168–179.

Mousty, P., & Leybaert, J. (1999). Evaluation des habilités de lecture et d'orthographe au moyen de BELEC: données longitudinales auprès d'enfants francophones testés en 2ème et 4ème années. *Revue Européenne de Psychologie Appliquée*, *49*, 325–342.

Mussolin, C., Mejias, S., & Noël, M. P. (2010). Symbolic and nonsymbolic number comparison in children with and without dyscalculia. *Cognition*, *115*(1), 10–25. https://doi.org/10.1016/j.cognition.2009.10.006

Pasquali, L., Wechsler, S. M., & Bensusan, E. (2002). Matrizes Progressivas do Raven Infantil: um estudo de validação para o Brasil [Raven's Colored Progressive Matrices for Children: a validation study for Brazil]. *Avaliação Psicológica*, *1*(2), 95–110.

Perfetti, C. A. (1985). *Reading ability*. Oxford University Press.

Piérart, B., & Grégoire, J. (2004). Déchiffrer et coprendre: le test de closure en lecture revisité ètalonnage belge du L3 de Lobrot. *Le Langage et l'Homme*, *39*(2), 87–100.

Pinheiro, A. M. V. (2013). *Prova de Leitura e de Escrita de palavras e de pseudopalavras* Relatório Relatório Técnico Final aprovado pela Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG (FAPEMIG). Número do processo: APQ-01914-09.

Pinheiro, A. M. V. (2015). Frequency of Occurrence of Words in Textbooks Exposed to Brazilian children in the Early Years of Elementary School. From *CHILDES – Child Language Data Exchange System*. http://childes.talkbank.org/derived

Pinheiro, Â. M. V., Vilhena, D. A., & Santos, M. A. C. (2017). PROLEC-T – Prova de Compreensão de Texto: Análise de suas Características Psicométricas [PROLEC-T - Text comprehension test: psychometric properties analysis]. *Temas em Psicologia [Trends in Psychology]*, *25*(3). http://dx.doi.org/10.9788/tp2017.3-08.

Reybroeck, M., & Hupet, M. (2009). Acquisition of number agreement: effects of processing demands. *Journal of Writing Research*, *1*(2), 153–172. https://doi.org/10.17239/jowr-2009.01.02.3

Rousselle, L., & Noël, M. P. (2007). Basic numerical skills in children with mathematics learning disabilities: A comparison of symbolic vs non-symbolic number magnitude processing. *Cognition*, *102*(3), 361–395. https://doi.org/10.1016/j.cognition.2006.01.005

Saadatnia, M., Ketabi, S., & Tavakoli, M. (2016). EFL Learners' Levels of Comprehension Across Text Structures: A Comparison of Literal and Inferential Comprehension of Descriptive and Enumerative Expository Texts. *Journal of Psycholinguist Research*, *45*(6), 1499-1513. https://doi.org/10.1007/s10936-016-9414-6

Salles, J. F., & Paula, F. V. (2016). Text reading comprehension and its relationship with executive functions. *Educar em Revista*, (62), 53-67. https://doi.org/10.1590/0104-4060.48332

Saur, A. M., & Loureiro, S. R. (2012). Psychometric properties of the Strengths and Difficulties Questionnaire: a literature review. *Estudos de Psicologia (Campinas), 29*(4), 619–629. https://doi.org/10.1590/S0103-166X2012000400016

Serniclaes, W., Heghe, S., Mousty, P., Carré, R., & Sprenger-Charolles, L. (2004). Allophonic mode of speech perception in dyslexia. *Journal of Experimental Child Psychology, 87*(4), 336–361. https://doi.org/10.1016/j.jecp.2004.02.0011

Stanovich, K. E., Cunningham, A. E., & Feeman, D. J. (1984). Intelligence, Cognitive Skills, and Early Reading Progress. *Reading Research Quarterly*, *19*(3), 278-303. https://doi.org/10.2307/747822

Sucena, A., & Castro, S. L. (2010). *Aprender a Ler e Avaliar a Leitura. O TIL: Teste de Idade de Leitura* (2nd ed.). Coimbra: Almedina.

Vilhena, D. A., & Pinheiro, A. M. V. (2016). Revised version of the Scale of Evaluation of Reading Competence by the Teacher: final validation and standardization. *Universitas Psychologica*, 15(4), 1-13. https://doi.org/10.11144/Javeriana.upsy15-4.efvs

Vilhena, D. A., Sucena, A., Castro, S. L., & Pinheiro, Â. M. V. (2016). Reading Test-Sentence Comprehension: An Adapted Version of Lobrot's Lecture 3 Test for Brazilian Portuguese. *Dyslexia*, *22*(1), 47-63. https://doi.org/10.1002/dys.1521