Psicologia: Teoria e Pesquisa
2018, v.34, e3437

DOI: https://dx.doi.org/10.1590/0102.3772e3437

School and Developmental Psychology

Word Decoding Task: Item Analysis by IRT and Within-Group Norms

Patrícia Silva Lúcio[1,*], Hugo Cogo Moreira[2], Adriana de Souza Batista Kida[2], Carolina Alvez Ferreira de Carvalho[2], Ângela Maria Vieira Pinheiro[3], Jair de Jesus Mari[2], & Clara Regina Brandão de Avila[2]

[1] *Universidade Estadual de Londrina, Londrina, PR, Brasil*

[2] *Universidade Federal de São Paulo, São Paulo, SP, Brasil*

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

ABSTRACT – This paper reports the performance of a representative sample of 747 students (52.5% female), from 2nd to 5th year of elementary education from private and public (83.8%) schools of Sao Paulo city. The children performed the Form A of Word Reading and Spelling Task (WRST) containing 48 low-frequency words presented in a card. Data were analyzed using models of Item Response Theory. We observed high levels of accuracy. The analysis selected 24 items, which presented low to moderate discrimination and difficulty indices. There were mean differences between grades, but not sex or school type. We report percentile norms for the grades for the WRST´S (Form) Reduced Version. The results support preceding studies with the word decoding tasks in Brazilian Portuguese, which attested to the quasi-regular character of that language.

KEYWORDS: decoding, Item Response Theory, Brazilian Portuguese, quasi-regular

Tarefa de Decodificação: Análise de Itens pela TRI e Normas Intragrupo

RESUMO – O estudo investiga o desempenho de uma amostra representativa de crianças (52,5% meninas) cursando do 2º ao 5º ano de escolas particulares e públicas (83,8%) de São Paulo (N = 747). As crianças leram em voz alta a Forma A da Prova de Leitura e de Escrita de Palavras (PLEP), constituída de 48 palavras de baixa frequência, apresentadas em um cartão. Os dados foram analisados utilizando modelos da Teoria de Resposta ao Item. Observaram-se altos índices de acurácia. Foram selecionados 24 itens, os quais apresentaram, em média, índices de discriminação e de dificuldade de baixos a moderados. Houve efeito de escolaridade, mas não de sexo ou tipo de escola. Normas em percentis foram reportadas para a versão reduzida da Forma A da PLEP. Os resultados condizem com estudos que utilizaram tarefas de decodificação de palavras no Brasil e atestam para o caráter quase-regular do idioma.

PALAVRAS-CHAVE: decodificação, Teoria de Resposta ao Item, português brasileiro, quase-regular

Decoding is defined by Perfetti and Hogaboam (1975) as the process of transferring the written code into the language code, what is followed by a vocalization of the decoded unit in a task of reading words aloud. In addition, decoding is related to the process of extracting semantic information of the words available in the mental lexicon (Stanovich, 1982). In those statements, is implicit the idea that decoding involves a process of learning. Therefore, learning to read in alphabetic languages is related to mastering the connections between the orthography and the phonology of the words.

How difficult it will be for the children to learn the connections between sound and speech heavily depends on the orthographic depth, the degree to which languages vary in terms of graphemes to phoneme consistencies (Katz & Frost, 1992). It means that the transparency level of the language will reverberate in learning to read, *i.e.*, will

* Email: pslucio@gmail.com

determine how the children will apprehend the language code. Consequently, to investigate word recognition accuracy at different spellings is important at least for two reasons. First, it allows verifying the current performance of the children (and therefore setting the standard parameters of learning to read). Second, it enables us to formulate theories about the development (or stages) of the process of reading (Wimmer & Goswami, 1994).

Some authors (*e.g.*, Seymour, Aro, & Erskine, 2003) argue that the orthographic depth differs between languages. Thus, language can range from deep orthography, with many inconsistencies and complexities (such as English), to shallowness, with relatively predictable pronunciations (as Finnish and Greek). The Portuguese language is classified by Seymour and colleagues (Seymour et al., 2003) as having an intermediate orthography depth, with simple structure (*i.e.*, the predominance of open consonant-vowel [CV] syllables). This view is compatible with that of Parente, Silveira, and Lecours (1997) who, analyzing the grapheme to phoneme correspondence in the Brazilian Portuguese, categorized it as an "almost regular" language. In fact, Brazilian Portuguese presents only two cases of inconsistencies in grapheme to phoneme correspondences[1] what makes reading isolated words aloud a relatively easy task for speakers of Portuguese, even for the beginning readers or the struggling ones.

The research of Lúcio, Moura, Nascimento, and Pinheiro (2012) supports that hypothesis. In this work, children from 2nd to 5th grades of Elementary School read a database containing 323 low frequency isolated words (Pinheiro, 2007), more than 50% of which were irregular. As results, only five items had accuracy (*i.e.*, proportion of correct responses) lower than 40%. Taking into account that these are the kinds of items that cause more difficulties for readers, these are striking results (*e.g.*, Coltheart, Rastle, Langdon, & Ziegler, 2001).

The quasi-regular nature of the grapheme-phoneme correspondences in Brazilian Portuguese language is probably in the core of the ceiling effects frequently found in researches that investigate the reading of single words aloud by children. The accuracy found studies with children from 2nd to 5th grades is around 70%-96% (*e.g.*, Capovilla & Capovilla, 1996; Godoy, Defior, & Pinheiro, 2007; Guimarães, 2004; Justi & Justi, 2009; Lúcio et al., 2012; Salles & Parente, 2007). Broadly, these studies include specific populations in their sample (such as dyslexics or poor readers).

This picture does not change for the only standardized measure of decoding words currently available in Brazil, the reading subtest of the Academic Performance Test (Stein, 1994). A study evidenced that such subtest exhibit a ceiling effect even between 2nd grade children (Lúcio, Pinheiro, & Nascimento, 2009), and a classical item analysis showed that only 38% of their items presents adequate D index (Lúcio & Pinheiro, 2014). In classical item analysis, the D index is given by the performance difference (mean correct responses in the item) between the highest scoring group of subjects (75th percentile or higher in the task) and the lowest scoring group (defined by the 25th percentile or lower). It produces a number between -1.0 and 1.0, so that an adequate D index is a number greater than 0.30 (Urbina, 2014). Additionally, in a literature review, Knijnik, Giacomonia, and Stein (2013) pointed out the need for new studies for updating the test. In line with that, some recent efforts have tried to promote new norms and psychometric inquiries for the test, with satisfactory results (Cogo-Moreira et al., 2013; Lúcio & Pinheiro, 2014; Knijnik, Giacomonia, Zanon, & Stein, 2014).

In Brazil, there is a lack of formal assessment instruments of the ability of reading words measuring the latent trait throughout the manifest/observable variables. At the same time, as pointed out, the unique standardized test actually available for evaluating reading ability presents psychometric issues. Furthermore, the relative shallowness of the language makes the reading task not very discriminative, which demonstrates the importance of psychometric studies for such a task. Thereby, the present study proposes to evaluate the performance of a representative sample of children from the 2nd to 5th year of elementary schools of Sao Paulo – Brazil, in a single-words aloud task. For this purpose, the Form A of the Word Reading and spelling task (WRST; *Prova de Leitura e Escrita de Palavras* [PLEP]) was used. It is an instrument for the evaluation of the reading and spelling abilities of school children (Pinheiro, 2013), which is being tested in many regions of Brazil. Our main goal is to investigate the appropriateness of this tool, followed by the adjustments that will be necessary in order to make it suitable for use by practitioners (in educational and clinical settings) and researchers in this field. As for results, two primary outcomes are expected: a) to confirm the findings of previous studies about the high accuracy of word reading in Brazilian Portuguese; and b) to provide evidences about the factorial structure of the WRST´S Form A and its fit indices; if the these indices converge for good fit indices, to provide norms to the resulting version of the instrument regarding Brazilian children attending the early years of schooling.

1 After the implementation of the Portuguese Language Orthographic Agreement, some new irregularities for reading are present (as for the word "linguiça" that used to receive the umlaut (¨) on the vowel <u>, what indicated that the digraph <gu> was pronounced as /gwi/ and not as /ʒi/). As there was not a word in the list that was affected by such changes, we kept the Parent's et al. classification for our purpose.

METHOD

Participants and Sampling

The sample was composed of 747 students from 2nd to 5th grade of Elementary Schools, boys and girls (52.5%) enrolled in private (16.2%) and in public (45.2% from state and 38.6% from municipal) schools from Sao Paulo city. Such proportionality was based on the Brazilian Scholar Census of 2012 (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP], 2013), and it was used to obtain a proportional representative scholar sampling from Sao Paulo's city. At random and following such proportionality, (using a stratified sampling strategy), 21 schools were selected.

Participants were indicated by their teachers according to the following criteria: no complaints or difficulties in learning how to read or write; no indication of school retention; absence of complaints of visual (uncorrected) and auditory sensory deficits, cognitive, behavioural and/or neurological disorders.

After identifying the eligible children by the teacher, the research team contacted the parents via a letter that described the study (its aims, procedures, and measurements), avoiding technical, scientific vocabulary. Together with the letter, it was requested the parents' written informed consent for their children's participation. The Ethical Committee from Federal Sao Paulo University approved this research (registration number: 38406/12).

As a final criterion of admission in the study, a screening task was used to identify and withdraw of the sample the struggling readers. We used this criterion because the same sample of children underwent reading comprehension tasks for another study; thereby a minimum proficiency in decoding was required for inclusion in the present study. For this purpose, it was used an in-text speed task. The children read a short text aloud to the examiner. It was used different texts (ranging from 206 and 235 words) for each school grade. After the children had finished reading the two first paragraphs, the examiner started clocking time. The rate of reading in the timeout of 60 seconds was the intake criterion. The cut-off was 50, 66, 77, and 95 read words for the 2nd to 5th year, respectively. The children who did not succeed the task were withdrawn from the study.

Materials

For evaluating reading ability, we used the WRST (Pinheiro, 2013). The task is composed of two sets of 48 low-frequency words common to children from the 2nd to the 5th grade (Form A and B, respectively). They were chosen according to the vocabulary extracted from published textbooks used in primary education in Brazil (Pinheiro, 1996, 2015). In each form of the task, the words vary in number of letters (4 to 8 letters) and in levels of phoneme-to-grapheme and grapheme-to-phoneme regularity. In the present study, we used only the Form A. The words were randomly presented in an A4 size paper, in capital letter, and in Arial 14.

Procedures

Trained speech-language therapists (n = 10) tested each child individually at appropriated rooms in their schools. The speech-language therapists were assigned by school. The children were asked to read aloud the items of the PLEP, which were presented on a card (font: Arial; size: 14). The accuracy was defined by the following criteria: words pronounced at once, without segmentations or self-correction, and accordingly the standard or local dialect, was considered correct. In such cases, children received a score of 1.0 point; otherwise, the answer was scored as 0.0 point.

Statistical Analysis

Firstly, two-parameter logistic models (2PL) were used in this study considering the weighted least squares means and variance adjusted (WLSMV) estimator (which uses a diagonal weight matrix with standard errors and mean- and variance- adjusted chi-square test statistics) and tetrachoric correlations. We choose a 2PL because is not reasonable to argue that all the items would have the same discrimination and all them equal to 1. Additionally, in an open-ended task, such as reading words aloud, is not necessary to build 3PL model, because no correct answers are expected by chance.

For the evaluation of model fit information, the following fit indices were considered: chi-square (χ^2), the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFT), the Tucker-Lewis Index (TLI), and the Weighted Root Mean Square Residual (WRMR) parameters. As criterion of goodness of fit, we used the following p-value of $\chi^2 \geq .05$; RMSEA < .06; CFI and TLI ≥ .95; and WRMR < 1.0 (Hoyle, 2012; Hu & Bentler, 1999; Loehlin, 2004).

Item selection pursues the same word list task for all grades. Differences between grades, type of school, and sex were tested. Levene's test for univariate ANOVA showed equality of error variance for type of school and sex ($F = 0.07$, $p = 0.934$ and $F = 3.275$, $p = 0.07$, respectively), but not for grades ($F = 0.536$, $p < .001$). For this reason, we used Kruskal-Wallis Test for general differences between grades and Dunnet's C as post hoc tests (the most appropriated test of pairwise comparisons for large samples when the assumption of equality of error variance is violated). For type of school and sex, t-tests for independent samples were conducted. The null hypothesis was rejected when $p < 0.05$. Statistical analysis were performed with Mplus 7.11 (Muthén & Muthén, 2012) and SPSS 20.0 for Windows (2011).

RESULTS

For a descriptive view, the proportion of correct answers to the 48 items of the PLEP is showed in Table 1, top (for the general sample and subsamples of grades, sex, and types of school). Consistent with previous studies with Brazilian Portuguese, the accuracy was relatively high, ranging from almost 64% of correct answers for the 2nd grade to 85% to the 5th grade (77% for the general sample).

Following the criteria described in the method, unidimensional models were obtained for the four grades, resulting in models with goodness of fit indices based on the item selection. The models were composed of 37 words for the 2nd grade group; 40 words for the 3rd grade group; 36 words for the 4th grade, and 37 words for the 5th grade group. From each of those four models, only the repeated items were kept (ex: *vasilha*; "canister"), because our objective was to adjust Pinheiro´s (2013) WRST (Form A) so that the reduced version would contain common words to the grades. Thereby, from the 48 original items, it remained 24 words from which is showed the descriptive statistics (Table 1, bottom) and was generated models for each grade (Table 2). As can be seen in Table 2, all the models presented goodness of fit based on the parameters cutoff mentioned in the method section. Only the p-value of the chi-squared statistics for the 2nd grade was bellow speciation, but as qui-squared is influenced by sample size, it is not taken as a definitive parameter (Hu & Bentler, 1999). Tables 3 and 4 presents discrimination and difficulty parameters. From this table, it is possible to see that both discrimination and difficulty index are low.

For the set of 24 items, Kruskal-Wallis Test showed mean differences between school grades (χ^2 = 134,836, df = 3, p < 0.001) and Dunnet's C post hoc test were unable to show differences only between the older children (4th and 5th grades). All the mean differences were in the expected direction, that is, older children with greater scores. There was a significant difference between the schools systems ($t(695) = 1.989$, $p = 0.047$), favorable to the children of private schools. Nevertheless, the value of the effect size was almost low ($d = 0.21$), which means that this statistical difference may not have a practical significance. In line with this hypothesis, is the fact that none of the grades presented statistical significant differences between the schools systems (2nd grade: ($t(177) = 1.701$, = p 0.091, $d = 0.37$; 3rd grade: $t(170) = 1.894$, $p = 0.060$, $d = 0.39$; 4th grade: $t(170) = 0.452$, $p = 0.652$, $d = 0.10$; 5th grade: $t(172) = 0.475$, $p = 0.635$, $d = 0.09$). Similarly, the differences regarding children's sex were not significant ($t(698) = 0.362$, $p = 0.717$, $d = 0.03$). Because there were not significant differences between type of school or sex, we present separated percentiles only for the four school grades under investigation (Table 5). Descriptive statistics of theta distribution of the sample is showed Table 6 and the distribution of Test Information Function is showed in Figure 1.

Table 1. *Proportion (%) of Correct Answers for the 48 Original Items (Top) and for the 24 Selected Items (Bottom)*

Grade	General Sample	Boys	Girls	Public Schools	Private Schools
Original items (48 items)					
2nd	63.9	62.1	65.7	62.4	72.7
3rd	76.1	76.7	75.9	75.0	82.5
4th	82.8	83.3	82.3	82.32	84.7
5th	85.1	85.7	84.6	84.7	87.2
All grades	76.8	76.3	77.2	75.9	81.5
Selected items (24 items)					
2nd	54.5	52.5	56.4	53.2	61.5
3rd	68.3	69.7	67.6	67.2	74.7
4th	76.9	77.9	76.0	76.6	78.6
5th	80.1	80.7	79.6	79.8	81.4
All grades	69.8	69.5	70.1	69.1	73.8

Table 2. *Model Fit Information for the Grades Groups in the 24 Selected Words*

Model Index	2nd	3rd	4th	5th
χ^2	290.209	270.471	251.995	260.710
df	252	252	252	252
p-value	0.049	0.202	0.489	0.340
RMSEA	0.028	0.020	0.000	0.014
90% C.I.	0.002-0.042	0.000-0.037	0.000-0.029	0.000-0.033
CFI	0.973	0.976	1.000	0.984
TLI	0.970	0.974	1.000	0.982
WRMR	0.863	0.869	0.771	0.840

Table 3. *Discrimination Parameter and Standard Errors (in Brackets) for the 24 Words of Pinheiro's (2013) WRST (Form A) Reduced Version [Versão reduzida da Forma A da PLEP (Pinheiro, 2013).*

Word	Translation	2nd	3rd	4th	5th
Farta	Satiate	0.831 (.17)	0.351 (.11)	1.126 (.27)	0.684 (.25)
Nublado	Cloudy	0.791 (.17)	0.481 (.12)	0.889 (.22)	0.723 (.19)
Treze	Thirteen	0.432 (.11)	0.295 (.10)	0.662 (.15)	0.324 (.12)
Enxuto	Wiped	0.504 (.11)	0.854 (.18)	0.740 (.14)	1.063 (.21)
Famoso	Famous	0.781 (.13)	0.767 (.17)	0.595 (.18)	0.627 (.15)
Cigana	Gipsy	0.473 (.11)	0.291 (.11)	0.550 (.13)	0.853 (.18)
Manhoso	Sly	1.041 (.17)	0.997 (.19)	1.141 (.25)	0.799 (.20)
Reflexo	Reflection	1.091 (.17)	0.643 (.14)	0.928 (.20)	0.720 (.20)
Universo	Universe	1.072 (.17)	0.979 (.18)	0.964 (.29)	0.803 (.23)
Doutora	Doctor	1.110 (.19)	0.510 (.13)	0.997 (.21)	0.815 (.27)
Bengala	Cane	0.645 (.12)	0.839 (.16)	0.907 (.18)	0.720 (.17)
Bexiga	Bladder	1.075 (.17)	0.635 (.14)	0.673 (.17)	1.307 (.27)
Picada	Sting	0.491 (.12)	0.750 (.15)	0.903 (.20)	0.828 (.20)
Mariposa	Moth	0.864 (.15)	0.668 (.14)	0.932 (.20)	0.818 (.17)
Frade	Shoveling	0.586 (.12)	0.543 (.13)	0.725 (.15)	0.631 (.15)
Redonda	Round	0.742 (.13)	0.589 (.13)	0.611 (.16)	0.210 (.13)
Faxina	Cleaning	0.726 (.13)	0.582 (.14)	0.414 (.12)	0.477 (.13)
Bondosa	Kind	0.878 (.15)	0.566 (.13)	0.645 (.14)	0.672 (.16)
Vasilha	Canister	0.640 (.15)	0.570 (.13)	0.869 (.15)	0.834 (.19)
Dengoso	Namby-pamby	0.631 (.12)	0.896 (.17)	0.580 (.16)	0.370 (.13)
Exposto	Exposed	0.556 (.12)	0.567 (.13)	0.842 (.16)	0.407 (.13)
Carroça	Wain	0.570 (.12)	0.696 (.14)	1.027 (.21)	0.535 (.14)
Formosa	Pretty	0.454 (.11)	0.315 (.11)	0.670 (.15)	0.394 (.12)
Bezerro	Calf	0.955 (.16)	0.886 (.17)	1.018 (.19)	0.602 (.15)

Table 4. *Difficulty Parameter and Standard Errors (in Brackets) for the 24 Words of Pinheiro's (2013) WRST (Form A) Reduced Version. [Versão reduzida da Forma A da PLEP (Pinheiro, 2013)].*

Word	Translation	2nd	3rd	4th	5th
Farta	Satiate	-1.029 (.20)	-2.060 (.52)	-1.800 (.32)	-1.987 (.43)
Nublado	Cloudy	-0.991 (.20)	-3.466 (.11)	-1.600 (.25)	-2.743 (.75)
Treze	Thirteen	-1.319 (.39)	-2.883 (1.0)	-1.339 (.29)	-2.681 (.98)
Enxuto	Wiped	0.873 (.26)	0.482 (.16)	-0.303 (.16)	-0.619 (.15)
Famoso	Famous	-0.509 (.16)	-1.402 (.28)	-2.682 (.69)	-2.098 (.45)
Cigana	Gipsy	0.061 (.21)	-0.817 (.44)	-0.676 (.23)	-0.883 (.19)
Manhoso	Sly	-0.200 (.13)	-0.974 (.18)	-1.277 (.20)	-1.821 (.36)
Reflexo	Reflection	0.168 (.12)	-1.176 (.27)	-1.720 (.28)	-2.193 (.48)
Universo	Universe	-0.343 (.13)	-1.192 (.20)	-2.203 (.43)	-2.545 (.54)
Doutora	Doctor	-0.230 (.12)	-2.357 (.56)	-1.512 (.24)	-2.311 (.53)
Bengala	Cane	-0.388 (.18)	-0.569 (.17)	-1.016 (.19)	-1.406 (.30)
Bexiga	Bladder	-0.217 (.13)	-0.860 (.23)	-1.961 (.42)	-1.335 (.19)
Picada	Sting	-1.086 (.31)	-1.489 (.29)	-1.530 (.26)	-1.869 (.35)
Mariposa	Moth	-0.120 (.14)	-0.512 (.19)	-1.020 (.20)	-1.017 (.21)
Frade	Shoveling	-0.296 (.19)	-0.465 (.22)	-0.972 (.21)	-1.113 (.28)
Redonda	Round	0.022 (.15)	-0.624 (.22)	-1.536 (.35)	-4.011 (2.37)
Faxina	Cleaning	-0.111 (.16)	-0.192 (.19)	-1.113 (.37)	-1.181 (.36)
Bondosa	Kind	-0.148 (.14)	-1.386 (.33)	-1.347 (.29)	-1.552 (.34)
Vasilha	Canister	1.282 (.29)	0.621 (.22)	0.041 (.14)	-0.294 (.15)
Dengoso	Namby-pamby	-0.110 (.17)	-1.070 (.20)	-1.873 (.48)	-2.148 (.75)
Exposto	Exposed	0.390 (.20)	-0.097 (.19)	-0.959 (.19)	-1.531 (.50)
Carroça	Wain	-0.506 (.23)	-1.143 (.25)	-1.255 (.20)	-1.965 (.49)
Formosa	Pretty	-0.395 (.23)	-1.138 (.49)	-0.907 (.24)	-2.251 (.69)
Bezerro	Calf	0.009 (.13)	-0.442 (.15)	-0.866 (.16)	-1.420 (.34)

Table 5. *Percentile norms for the 24 selected items (raw scores)*

Percentiles	2nd	3rd	4th	5th
5	3.00	9.00	9.00	12.00
25	9.00	13.00	16.00	17.00
50	13.00	17.00	20.00	20.00
75	18.00	20.00	22.00	23.00
95	22.00	23.00	24.00	24.00

Table 6. *Descriptive Statistics for Theta Distribution*

Grade	Minimum	Maximum	Mean	S.D.
2nd	-1.502	1.316	-0.009	0.589
3rd	-0.871	0.766	-0.003	0.387
4th	-1.644	0.867	-0.032	0.566
5th	-1.238	0.726	-0.022	0.497

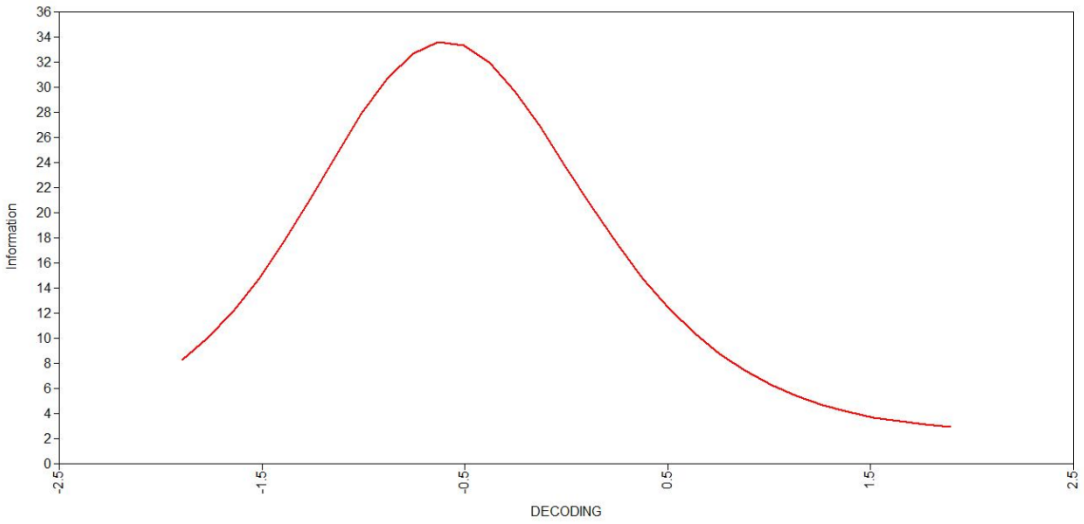

Figure 1. Information Test Function for all grades

DISCUSSION

With the main goal of reducing the lack of formal assessment tools for reading words aloud in Brazilian Portuguese, this study reported IRT analysis for a set of items extracted from Pinheiro's (2013) WRST (Form A). From this study, it was possible to generate a reduced version, of the instrument based on the fit index and to offer within-group norms for a representative sample of children from 2nd to 5th grades of Sao Paulo – Brazil. From the original set of 48 low frequency words, it was selected 24 items with goodness of fit for the four grades under study. In general, the results of the descriptive statistics for both set of 48 and 24 items are consistent with previous studies with Brazilian Portuguese speakers children (*e.g.*, Lúcio et al., 2012, 2009), which concluded that reading single words aloud is not too difficult, because of the quasi-regular character of the language (Parente et al., 1997). Indeed, in the original instrument used in this study, accuracy ranged from 64% to 85% (in the 2nd and 5th year, respectively), and this picture has not changed much after item selection, with accuracy 7% lower for the general sample. It is worth pointing out that the task was composed of low frequency words, much of which irregular ones. Additionally, we used a conservative criterion for accuracy in reading, in which syllabication followed or not by self-correction, was taken as an error. Therefore, it would be expected the accuracy reduction due to the cumulative factors (*i.e.*, low frequency and word's irregular structure, added to a conservative criterion for accuracy).

However, such phenomenon was not observed, and high levels of accuracy were maintained even between the youngest readers. This fact is in line with the hypothesis regarding the almost shallow nature of Brazilian Portuguese that could per se help the children in the process of decoding. Similar phenomenon occurs, for example, in German language (Wimmer & Goswami, 1994). The more predictable the pronunciation of the item, easier is decoding and consequently, higher the accuracy. The same can be stated about the IRT's difficulty parameter obtained for the items, most of which were easy, as in previous studies (*e.g.*, Cogo-Moreira et al., 2013; Knijnik et al.,

2014). For example, considering all grades, the highest difficulty's parameter (threshold) was 1.3 for the word *vasilha* ("canister") observed in the second year. Also, the lowest value was -4.0 observed in the fourth grade (*redonda* – "round"). It means that, in general, the selected items are easy and that even individuals with low theta levels are unlikely to fail in them.

The results for difficulty parameters could be in part expected because of the selection criteria adopted, that precluded completing the latent space defined for children of interest (*i.e.*, children with poor word-level reading skills). In fact, an examination of the theta distribution (Table 6) make clear that very low values of decoding ability ($\theta < -2.0$) were not sampled in this study (the lowest value was -1.7 in the 4th grade). Thus, the list of words suits more for children with moderate trait level, as shown in the Test Information Function (Figure 1). It means that more reliable results are found to this trait level.

It might be pointed out that the results obtained may not be explained by an inappropriate, non-representative sample of test items for the decoding domain. It is well known that ability estimates, unlike the ability scores, are "item-free" when a set of representative items are sampled in the study (Hambleton & Swaminathan, 1985). As our results are in line with previous studies using low frequency words as test items, and average children as sample, and those studies have showed high general levels of accuracy (*e.g.*, Justi & Justi, 2009; Pinheiro, 1995; Salles & Parente, 2007) and low discrimination D index for items (Lúcio et al., 2012), it is plausible to suppose these results are not due to sampling test items. There are not children with low levels of theta (*i.e.*, below than -2.0), as it might be noted in Table 6, due to the study selection criteria conditioned by the screening task (see Method).

But, why do high theta levels are also not sampled in this study (as can also be seen in Table 6)? If the screening task has precluded struggler's readers from taking part in the study and being the sample representative of the population of 2nd to 5th grades of Sao Paulo, it would be expected that the test would detect the best decoders, the ones who present the higher theta levels ($\theta > 2.0$). Again, the shallowness nature of the Brazilian Portuguese language is probably in the core of the explanation. It seems that the reading words aloud task is not hard enough to represent high levels of the decoding ability. For example, Cogo-Moreira and colleagues (Cogo-Moreira et al., 2013) showed similar results presented in this research, also with a randomized and representative sample of children and with low frequency words, mostly. The values obtained for the estimated *b* parameter was not higher than 0.062, and it is well-acknowledged that theta estimates surround the value of *b* (Embretson & Reise, 2000). A similar pattern of results was obtained by Knijnik et al. (2014), who found 0.290 as the highest value of *b*, using a sample of 1850 children from four Brazilian States. Therefore, it is very likely that, at least in Brazilian Portuguese, higher levels of theta will not be found in the task of reading words aloud.

For the majority of the items, parameters' estimates did not show satisfactory results for discrimination, which means that the items are not very precise at differentiating respondents around the location point. In other words, the low discrimination index obtained indicate that, in general, changes in the trait level produced small impact on the probability of success in the items (Embretson & Reise, 2000). Considering the Baker's cutoff for discrimination parameter classification (Baker, 2001), 40% of the items had low or very low indices of discrimination (minimum of 21% in the 4th year and maximum of 54% in the 2nd year). The best outcomes were obtained from the 4th grade, with almost 80% of the items presenting moderate discriminations. None of the items showed high indices of discrimination.

The analysis showed significant differences of means between grades, but not for type of school or sex. As expected, accuracy for older children was higher than those from earlier grades. Because this, we presented percentile norms separated for grades, but not for the other groups (Table 5). We expect that these norms are useful for both practitioners and researchers who evaluate the reading skills of children of the population investigated in this study.

Finally, we stress the relevance of this research to both the area of psychometrics and cognitive reading evaluation in Brazil. For our knowledge, only two published studies used Item Response Theory to evaluate items proprieties of reading single words aloud (Cogo-Moreira et al., 2013; Knijnik et al., 2014) and it is the first to make the analysis by grade. Additionally, the norms reported are useful for evaluating the reading skills of children from 2nd to 5th grades. Nevertheless, the function information curve show higher precision (amount of information) for children with average decoding abilities (theta levels between -1 and +1). As limitations, we point out the lack of struggling readers in the sample, and their inclusion must be done in future researches.

CONCLUSION

As hypothesized, the psychometric investigation carried out in the present study confirmed previous findings of accuracy in reading single words aloud in Brazilian Portuguese and pointed out to the almost shallowness of that language. Therefore, "near shallowness" character of Brazilian Portuguese might be in the core of the psychometric index obtained for the list of single words. Despite this, and although improvements in the discrimination index are required for the selected items, it can be concluded that the percentiles norms reported in this study for the Pinheiro's (2013) WRST (Form A) Reduced Version are adequate for assessing the reading skills of the population investigated (*i.e.*, children from 2nd to 5th year of elementary schools with normal levels of decoding ability). Therefore, further studies including struggling readers in the sample should be conducted in order to increase the latent space of the population of interest (although by inference, it can be assumed that struggling readers would reach the lowest levels of reading ability based on the given norms). Nevertheless, the quite goodness of fit index obtained for the final set of items of the instrument show that the provided parameters (percentile norms) may be useful for evaluating the ability of reading single words among school-aged children.

REFERENCES

Baker, F. B. (2001). *The basics of item response theory*. Washington: ERIC Clearinghouse on Assessment and Evaluation.

Capovilla, F. C., & Capovilla, A. G. S. (1996). Leitura, ditado e manipulação fonêmica em função de variáveis psicolinguísticas em escolares de terceira a quinta série com dificuldades de aprendizagem. *Revista Brasileira de Educação Especial*, *2*(4), 53–71. Retrieved from http://educa.fcc.org.br/scielo.php?script=sci_arttext&pid=S1413-65381996000100006&lng=pt&nrm=iso

Cogo-Moreira, H., Carvalho, C. A. F., Kida, A. de S. B., Avila, C. R. B. de, Salum, G. A., Moriyama, T. S., … Mari, J. de J. (2013). Latent class analysis of reading, decoding, and writing performance using the Academic Performance Test: Concurrent and discriminating validity. *Neuropsychiatric Disease and Treatment*, *9*, 1175–85. doi:10.2147/NDT.S45785

Coltheart, M., Rastle, K., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256. doi:10.1037/0033-295X.108.1.204

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: L. Erlbaum Associates.

Godoy, D. M. A., Defior, S., & Pinheiro, Â. M. V. (2007). Impacto do método de alfabetização sobre o desenvolvimento da consciência fonêmica, da leitura e da escrita no português do Brasil. *Educação: Temas e Problemas*, *4*, 81–99.

Guimarães, S. R. K. (2004). O papel das pistas do contexto verbal no reconhecimento de palavras. *Psicologia Em Estudo*, *9*(2), 279–289. doi:10.1590/S1413-73722004000200014

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and appliccations*. New York: Springer Science+Business Media, LLC.

Hoyle, H. H. (2012). *Handbook of structural equation modeling*. New York: The Guilford Press.

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

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2013). *Censo da educação básica: 2012 – Resumo técnico*. Brasília. Retrieved from http://download.inep.gov.br/educacao_basica/censo_escolar/resumos_tecnicos/resumo_tecnico_censo_educacao_basica_2012.pdf

Justi, C. G. N., & Justi, F. R. dos R. (2009). Os efeitos de lexicalidade, freqüência e regularidade na leitura de crianças falantes do português brasileiro. *Psicologia: Reflexão e Crítica*, *22*(2), 163–172. doi:10.1590/S0102-79722009000200001

Katz, L., & Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. In R. Frost & L. Katz (Eds.), *Advances in Psychology 94: Orthography, Phonology, Morphology, and Meaning* (pp. 67–84). Amsterdam: North Holland Elsevier Science Publisher.

Knijnik, L. F., Giacomoni, C., & Stein, L. M. (2013). Teste de desempenho escolar: Um estudo de levantamento. *Psico USF, 18*(3), 407-416. doi: 10.1590/S1413-82712013000300007

Knijnik, L. F., Giacomonia, C. H., Zanon, C., & Stein, L. M. (2014). Avaliação dos subtestes de leitura e escrita do teste de desempenho escolar através da Teoria de Resposta ao Item. *Psychology, 27*(3), 481-490. doi: 10.1590/1678-7153.201427308

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). New Jersey: Lawrence Erlbaum Associates.

Lúcio, P. S., Moura, R. J. de, Nascimento, E. do, & Pinheiro, Â. M. V. (2012). Construção de uma tarefa de leitura em voz alta de palavras: Análise psicométrica dos itens. *Psicologia: Reflexão e Crítica*, *25*(4), 662–670. doi:10.1590/S0102-79722012000400005

Lúcio, P. S., & Pinheiro, Â. M. V. (2014). Novos estudos psicométricos para o subteste de leitura do teste de desempenho escolar. *Temas em Psicologia*, *22*(1), 109–119. doi:10.9788/TP2014.1-09

Lúcio, P. S., Pinheiro, Â. M. V., & Nascimento, E. do. (2009). O impacto da mudança no critério de acerto na distribuição dos escores do subteste de leitura do teste de desempenho escolar. *Psicologia em Estudo, 14*(3), 593–601. doi:10.1590/S1413-73722009000300021

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide* (7th ed.). Los Angeles: Muthén & Muthén.

Parente, M. A. de M., Silveira, A. da, & Lecours, A. R. (1997). As palavras do português escrito. In *Dislexia: Implicações do sistema de escrita do português* (pp. 41–55). Porto Alegre: Artes Médicas.

Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology*, *67*(4), 461–469. doi:10.1037/h0077013.

Pinheiro, Â. M. V. (1995). Reading and spelling development in Brazilian Portuguese. *Reading and Writing*, *7*(1),111-138.

Pinheiro, Â. M. V. (1996). *Contagem de frequência de ocorrência de palavras expostas a crianças na faixa pré-escolar e séries iniciais do 1º grau*. São Paulo: Associação Brasileira de Dislexia.

Pinheiro, Â. M. V. (2007). Lista de palavras. In I. Sim-Sim & F. L. Viana (Eds.), *Para a avaliação do desempenho de leitura* (pp. 120–130). Lisboa: Gabinete de Estatística e Planejamento da Educação (GEPE).

Pinheiro, Â. M. V. (2011). Transparência ortográfica e o efeito de retroalimentação fonológico grafêmica: implicações para a construção de provas de reconhecimento de palavras. In L. M. Alves, R. Mousinho, & S. A. Capellini (Eds.), *Dislexia: Novos temas, novas perspectivas* (pp. 131–146). Rio de Janeiro: Wak.

Pinheiro, Â. M. V. (2013). *Validação e estabelecimento de normas de uma prova computadorizada de reconhecimento de palavras para crianças* (Relatório técnico final aprovado pela FAPEMIG em novembro de 2013. 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. Childes - Child Language Data Exchange System*. Retrieved from http://childes.talkbank.org/derived

Salles, J. F. de, & Parente, M. A. de M. (2007). Avaliação da leitura e escrita de palavras em crianças de 2ª série: Abordagem neuropsicológica cognitiva. *Psicologia: Reflexão e Crítica*, *20*(2), 220–228. doi:10.1590/S0102-79722007000200007

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology (London, England: 1953)*, *94*(2), 143–74. doi:10.1348/000712603321661859

SPSS, I. (2011). *20.0 for Windows*. [Computer software]. Chicago, IL: SPSS Inc.

Stanovich, K. E. (1982). Individual differences in the cognitive process of reading: I. word decoding. *Journal of Learning Disabilities*, *15*(8), 485–493. doi:10.1177/002221948201500809

Stein, L. M. (1994). *Teste de desempenho escolar*. São Paulo: Casa do Psicólogo.

Urbina, S. (2014). *Essentials of psychological testing*. Hoboken, New Jersey: Wiley & Sons, Inc.

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition*, *51*(1), 91–103. doi:10.1016/0010-0277(94)90010-8

Submitted: 03/08/2015
Reviewed: 18/09/2016
Acceptted: 05/09/2018