Evaluation of the Computer-based Battery for Oral Language (BILO*v3*) through the Rasch model for a Brazilian Sample*

Evaluación de la batería computarizada para lenguaje oral (BILO*v3*) a través del modelo de Rasch para una muestra brasileña

Recibido: marzo 29 de 2013 | Revisado: mayo 17 de 2014 | Aceptado: mayo 17 de 2014

MARIA CRISTINA RODRIGUES AZEVEDO JOLY **
Universidade de Brasília, Distrito Federal, Brasil
ANELISE SILVA DIAS ***
Universidade Paulista, São Paulo, Brasil
CAROLINE TOZZI REPPOLD ****
Universidade Federal de Ciências da Saúde
de Porto Alegre, Brasil

ABSTRACT

There are few tests in Brazil to assess the comprehension of oral language, especially the ones that are computer-based. Thus, this study aimed at investigating psychometric characteristics for a computer-based battery for Oral Language, Bilo*v3* using the Rasch Model. 569 children collectively answered BILO*v3*, which consists of 6 tests ok. A general analysis of the battery indicated that the test *Completing Sentences* was the one with the highest index of difficulty, whereas the test *Completing* Stories was the lowest. However, in the latter, there are difficult items regarding Morpho-syntactic Comprehension, Logical-Verbal Organization and Story Interpretation. As far as precision was concerned, the Kuder-Richardson coefficient varied from 0.80 to 0.97, which shows that the test is consistent. These results are evidences of validity and precision for BILO*v3.*

Keywords
Oral language; rasch; irt; validity; precision; children evaluation

RESUMEN

En Brasil, existen pocas pruebas para evaluar la comprensión del lenguaje oral, en particular de las informatizadas. Por lo tanto, este estudio tuvo como objetivo investigar las características psicométricas de una batería computarizada, constituida por seis pruebas, para evaluar lenguaje oral, Bilo*v3*, utilizando el modelo de Rasch. En forma colectiva, 569 niños respondieron BILO*v3*. Un análisis general de la batería indica que la prueba para completar oraciones fue la de mayor índice de dificultad, mientras que las pruebas de completar historias obtuvo el más bajo, si bien, en esta última, se encuentran dificultades en cuanto a comprensión morfosintáctica, organización lógico-verbal e interpretación. En lo que se refiere a precisión, el coeficiente de Kuder-Richardson varió desde 0.8 hasta 0.97, demostrándose así la consistencia de la prueba. Estos resultados son evidencias de validez y precisión para BILO*v3*.

Palabras clave
Lenguaje oral; rasch; irt; validez; precisión; evaluación niños

doi:10.11144/Javeriana.UPSY13-4.ecbb

Para citar este artículo: Joly, M. C. R. A., Dias, A. S., & Reppold, C. T. (2014). Evaluation of the computer-based battery for oral language (BILOv3) through the Rasch model for a Brazilian sample. *Univeristas Psychologica, 13*(4), 1367-1378. http://dx.doi.org/10.11144/Javeriana.UPSY13-4.ecbb

[*] Research paper with the financial support by CNPq FAPESP, Brasil

[**] Universidade de Brasília, Distrito Federal, Brasil E-mail: joly@unb.br

[***] Universidade Paulista, São Paulo, Brasil E-mail: anelisesd@gmail.com

[****] Programa de Pós-Graduação Stricto Sensu em Ciências da Saúde e Ciências de Reabilitação/ Universidade Federal de Ciência da Saúde de Porto Alegre, Porto Alegre, Brasil

E-mail: carolinereppold@yahoo.com.br

language is a means of representation and communication that involves intrinsic meanings and, therefore, must be understood in the context of social interaction. (Araújo et al., 2010). In order to establish proper communication, the subject must be able to use a linguistic variety of styles or registers that vary according to contingencies and the standards of interaction learned. Particularly for children, the use of language, the production and comprehension of language are relevant indicators of the access they have to learning situations and group affiliation (Wagner et al., 2010).

According to Ferreira et al. (2010) and Sternberg (2010), language is divided in two different areas: receptive and expressive language. Receptive language is the possibility to understand words and gestures, while expressive language is the ability to use gestures, words, signs and other written symbols for communication. Receptive language is associated with the semantic sphere and the comprehension of the linguistic code, and it allows the inference of meaning from plausible contexts (Johnson et al., 2009). On the other hand, expressive language is associated with the phonetic sphere of language and the ability to talk and communicate orally (Smeekens, Riksen-Walraven, & Bakel, 2008).

Traditionally, the role of oral language comprehension has received little scientific attention; more emphasis has been given to studies about oral production. However, the comprehension skills offer the base to learn new meanings and symbols, and both the comprehension of spoken words – receptive vocabulary, as well as sentences (receptive grammar) may develop at a point when the child still does not have the ability of oral production. As for children who are unable to talk, the comprehension of spoken language contributes for communicative exchange and provides the basis for their learning about the world as a whole (Geytenbeek et al., 2010).

In neuropsychological terms, language comprehension is increasingly seen as a process of wide distribution, including cortical and subcortical regions that go beyond temporal and parietal regions of the brain (Tremblay & Small, 2011). There is evidence of the relation between the brain development, language and other cognitive functions involved in the organization of linguistic sub-systems, such as phonology, syntax, semantics and pragmatics. Such relationships involve various mental processes that enable the understanding of a message and precede the act of speaking. More specifically, they refer to the functioning of a speaking person in activities of perception, comprehension, memory, production and meta-language (Crespo-Allende & Alfaro-Faccio, 2010; Gahyva & Hage, 2010; Hincapié-Henao et al., 2008; Rodríguez, Santana, & Caballero, 2012).

DSM IV-TR (American Psychiatric Association, 2002) classifies the Receptive Language Disorder as a condition in which language comprehension by the child is below the expected standards. Characteristics of this disorder vary according to the degree of the condition and the age of the child, and include limited speech, limited vocabulary, difficulty learning new words, errors in lexical choice, abbreviated sentences, simple grammar structures, limited variety of grammar and/or types of sentences, omission of critical parts of a sentence, unusual word order and slow language development in general. Children with this kind of disorder frequently start to speak at a later age and go through the several stages of expressive language development more slowly than expected.

Despite the relevance of the evaluation, alterations in receptive language for children in school age still has a very poor prognosis and there is the need for more research on efficient intervention and evaluation instruments (Bishop et al., 2006; Crespo-Allende & Alfaro-Faccio, 2010). In Brazil, there are few formal instruments available for clinical evaluation, particularly regarding the standardization of language comprehension patterns (Macedo et al., 2007), that enable the observation and description of processes directly involved with language, as well as the interaction among such factors and the related cognitive processes (Gahyva & Hage, 2010).

Most of the tests used in research that evaluates language, focus on receptive vocabulary and grammar comprehension. There is, however, the need to develop more work in ideal research situations that include oral comprehension of other areas that compose language, such as sequence comprehension of oral language (Bermúdez-James & Satre-Gómez, 2010; Gurgel et al., 2010).

The difficulty in using adequate psychometric instruments is even greater when computer-based tools are involved, given the lack of these in Brazil. When planning and developing computer-based tools installed in local computers or internet-based tools (Prieto, 1999), one must consider the possibility of using multimedia resources associated to programming techniques, which makes it viable to measure results or abilities that, until then, could not be measured (Bennett, 2001; Joly & Reppold, 2010; Wall, 2000).

According to Evers, Sijtsma, Lucassem, and Meijer (2010), the criteria used to evaluate the quality of tests, paper-based or computer-based, are especially focused on the analysis of theoretical references, of testing materials, of manuals and of accuracy and validity evidences. In both contexts, advanced statistical methods must be used to analyze items, such as the Item Response Theory (IRT), for the constant update of references (domain-referenced and criterion-referenced interpretation) and to estimate the accuracy of items, all of these considered according to the characteristics of the investigated sample.

One of the computer-based tests currently being studied in Brazil is the Computer-based Battery for Oral Language-BILO*v3* – BILO. This instrument, developed by Joly (2008), aims at evaluating oral language in children from five to ten years of age. There are several studies reporting evidences of content validity, results, criteria and precision. Since its creation, the battery has already been organized in three different versions due to technical adjustments that were needed in order to increase the appropriateness and psychometric qualities of tests that compose the battery. Considering the need to investigate the BILOv3 with samples from different parts of Brazil, the objective of this study is to present a study about validity evidences of the Computer-based Battery for Oral Language-BILO*v3* through the Rasch model for children from Porto Alegre (RS, Brazil).

Previous Studies with BILO

Studies conducted with BILO*v1* (Joly et al., 2008) were conducted with children and specialists in Psychology, Information Technology and Communication, who validated the battery of tests using the validation criteria of SAPI, a computer-based psychological evaluation system (Joly et al., 2005). TESTFACT was used for the analysis of each test in order to study the validity of results of this version. Results indicated that four of them attended the one-dimensional model - Morpho-syntactic Comprehension (MC), Logical Sequence (LS), Logical-Verbal Organization (LVO) and Completing Sentences (CSe) – while the others - Interpreting Stories (ISt), Completing Paragraphs (CP) and Completing Stories (CSt) still needed that items scoring below 0.30 be removed to attend such criteria.

The analysis of each item in BILO*v1* was made by testing each one of them through the Item Response Theory (IRT) using a three-parameter model, as well as evaluating the precision of each one of them. Alterations were suggested for MC, LVO, CSe, and CSt for items whose difficulty level (*b*) was too high; the tests LS, ISt and CP showed good adequacy to the three parameters. The precision measured by the Kuder Richardson Test (KR-20) varied from 0.64 (LVO) to 0.97(CSt). Analysis relating each test to the total score of BILO*v1* indicated that there is a connection between cognitive abilities and language abilities related to oral comprehension, validating the work and content of BILO*v1*.

Other evidences of validity for BILO*v1* were revealed by researches developed by Joly and Piovezan (2008) regarding mental maturity, by Almeida and Joly (2008) regarding vocabulary (performance in the Peabody Test of Vocabulary through Images) and by Issa (2008) regarding attention and hyperactivity. Based on these studies, it could be concluded that there are statistically significant differences for participants' performance according to their grade level and age. The study of Issa (2008) showed that there are no significant differences in score regarding participants' gender.

BILO*v2* was studied by Joly and Dias (2009), Joly, Piovesan et al. (2009) and Joly, Reppold, and Dias (2010). Such studies indicated evidences of precision for all tests, excepting Completing Paragraphs (CP), of criterion validity of BILOv2 regarding development and learning, and result validity of BILO*v2* through the correlation of tests with converging results.

BILO*v3* was developed taking into consideration the need to remove test CP – Completing Paragraphs, to reformulate items that presented vocabulary problems and to reduce the number of items so that all tests had the same number of questions. Validity of BILO version 3 was shown in studies of Joly (2010), Istome (2010), Soares et al. (2010) and Freitas (2011), conducted in different Brazilian cities (São Paulo, Itatiba, Macapá, and Natal), and by a correlation study with various tests in the expected magnitude and direction. Although the three versions of BILO have psychometric studies, tests still need to be analyzed through the Item Response Theory (IRT) in order to validate the battery, taking into account the relationship between the child's abilities and the item difficulty (Embretson & Reise, 2000; Tristán & Vidal, 2007; Van der Linden & Hamblenton, 2010).

Models of Item Response Test aim at evaluating the latent traits as non-observable abilities underlying the observable behavior analyzed by the answers in items of the test. Two characteristics of IRT can be mentioned. The first is about the possibility of attributing to the same metric scale the difficulty of the items, as well as the subjects' abilities. The mathematical model on which IRT is based allows the prediction of the probability of a person's right answers in a given ability represented by the results of the test (Fletcher, 1994); The second is in relation to the premise of providing invariable measures for cognitive performance, which does not depend on the items that compose the test or people tested. Thus, an IRT allows us to compare results of tests of variable difficulty and include results in cognitive performance in the same scale (Urbina, 2007).

This analysis model through the Item Response Test has been used in psychological and educational evaluation qualitative processes, in measuring instruments, such as the ability scales, to evaluate and follow the knowledge acquired by students, as well as the development of basic abilities, as it can help one reflect on items that are not working as expected. The observation of people's adaptation allows the identification of those who did not respond the test as expected. It is believed that these indicators can guide the process of measuring as they detect deficiency in adjusting or results that are too good to be true. In both cases further investigation is necessary in order to understand what happens to data (Wright and Stone, 1999), the object of investigation for BILOv3 by using the Rasch Model, aiming at identifying validity evidences for students from Porto Alegre (RS, Brazil).

Method

Sample

Subjects of this study were 569 children, ages five to 14 (M = 8.71; SD =1.4), regularly enrolled in schools. Subjects were from public and private schools in the city of Porto Alegre, state of Rio Grande do Sul, Brazil. Sample size was calculated according to "items/subjects ratio" (Pasquali, 1999). This criterion suggests a number of subjects ten times larger than the number of items when an exploratory analysis is to be made, such as in this study.

The education level of the tested children ranged from the first to the fourth year of basic education as defined by the new nine-year Brazilian education system. However, during data collection period, some 4th grade classes from the previous system still remained (3.3%). These classes were also included in the sample because children were in the studied age bracket. Gender distribution was uniform, with 52.5% girls, 45.7% boys, and 1.8% not identified.

Material

Computer-based Battery for Oral Language-BILOv3 (Joly, 2008)

The purpose of BILO is to assess oral comprehension by students within the first nine years of formal education under the present Brazilian education system. BILO*v3* was developed with the *Run Revolution* software, which offers multimedia resources. The program also interfaces with a MySQL database, for storage of test responses. It generates an application, which is installed in each computer. BILO consists of five tests, covering morpho-syntactic analysis, logical sequence, sentence comprehension, and story comprehension. Tests relating to language comprehension, in terms of its structuring into sentences and storylines, were prepared using the Cloze Oriented System – SOC (Joly, 2007). The SOC enables users to organize a text to be used in reading comprehension assessments, built according to specific criteria regarding number of words, omitted words, gap sizes and answer options, aiming to define different levels of comprehension difficulty. Because this assessment was focused on spoken language, with subjects that had not acquired the formal reading code, it was based on spoken instructions and responses presented in the form of simple line drawings, with no details. Drawings were selected based on symbols, meanings and contexts that were familiar to children. Each screen contains only one item of the test; after a response is chosen, respondent is asked to confirm his or her answer in order to continue the test. Responses and total time spent per item are recorded in a database.

Tests were applied collectively, guided by a test administrator and an assistant, and were administered to groups of 15 respondents at most, in a computer laboratory. The average application time was 40 minutes. Per item, per test, and general correction criteria were applied. Each test contains 10 multiple-choice items with three answer options. The contents of two options, including the correct answer, belong to the same grammatical class or category. The third option belongs to a different grammatical class or category than the correct answer. The total sum of points per item determines the score per test, up to a maximum score of 20 points. The total sum of scores per test results in the total score. The battery begins with an interactive tutorial to demonstrate the computer interface features to be used to answer the BILO. In this tutorial, respondent is asked to complete a few items, in order to practice the necessary abilities. Each test can be described as follows:

Test 1: Morpho-syntactic Comprehension (MC): aims to assess the relationship between a word and its graphical representation, revealing the comprehension of meaning.

Test 2: Logical Sequence (LS): assesses respondent's logical and temporal organization of visual stimuli organized into scenes that tell a story when arranged into the correct sequence.

Test 3: Logical-Verbal Organization (LVO): assesses receptive comprehension of the content of a story presented in its entirety both on video and orally, through logical organization of scenes that represent that content.

Test 4: Story Interpretation (SI): assesses receptive comprehension of three stories, through multiple-choice questions.

Test 5: Completing Sentences (CSe): assesses comprehension of words organized into sentences, where one noun has been omitted.

Test 6 - Completing Stories (CSt): assesses oral comprehension of stories. Each complete story is first introduced as a video (images and audio), and then by writing on a screen, accompanied by the respective audio, with response options for each item (images) omitted according to the SOC.

Procedure

The Institutional Ethics Committees of both universities that were involved in its execution authorized this study. Tests were applied collectively to groups of five to 10 respondents, in single sessions, guided by a test administrator and an assistant, in the computer laboratories of each school. Children answered the battery on individual computers, with headphones plugged in, for an average duration of 30 minutes.

Results and Discussion

Items of the Computer-based Battery for Oral Language – version 3 were analyzed using the Rasch model, since according to Evers, Sijtsma, Lucassem, and Meijer (2010), advances in this field require not only the construction of high-quality tests, but also the use of advanced statistical models, such as the Rasch, in addition to computer-based tests (Bennett, 2001; Wall, 2000). Each test was analyzed independently, observing the unidimensionality assumption (Embretson & Reise, 2000; Pasquali, 2007), as already applied in previous versions (Joly, 2009; Joly & Dias, 2009; Joly et al., 2008).

In the Morpho-syntactic Comprehension test (MC), based on results described in Table 1, items presented difficulty levels ranging from -0.76 to 0.62. This reveals that items were easy to answer, as the metric for this parameter ranges from -3 to +3, with the correct answer targeted at + 0.5 according to the estimated ability level of the subject (Linacre, 2002; Pasquali, 2007). It was observed that items could be grouped into three easier items (items 4, 7 and 9), four midrange items (1, 6, 8 and 10), and three items with difficulty values higher than 0.50 (2, 3 and 5).

Regarding infit and outfit values, the mean infit value for this test was 1.01 (*SD* = 0.18) and every item presented values within the critical limit range, namely 0.7 to 1.5, which is considered to be the criterion for good fit (Linacre, 2002). Given that infit refers to response patterns that do not fit the expectations for inliers (i.e., the subjects to whose ability levels the difficulty of the items was targeted), results of the morpho-syntactic comprehension test for the sample indicated that the group of items did not reveal discrepancies in the difficulty values of each item when targeted on the estimated abilities of subjects.

For outfit values, pertaining to unexpected response patterns by subjects with ability levels far from the difficulty of the items, the mean value for MC test items was 0.93 (*SD* = 0.32). Table 1 shows that only one item (9) presented values below the lower limit of the acceptable range (0.5), indicating that the answers given to this item did not fit the expected patterns based on the estimated MC abilities of respondents of this study.

The observed indices of correlation between a subject's MC ability and his or her response ranged from 0.34 to 0.59. According to Wright and Stone (2004), this variation in results is in line with expectations.

As regards the difficulty of the Logical Sequence test (LS), it was observed that the items presented difficulty values ranging from -0.48 to 0.76, revealing that items are easy to answer, under the metric described by Linacre (2002) and Pasquali (2007). Three of the items were easier (items 1, 4 and 3), five were of midrange difficulty (2, 5, 6, 9 and 10) and two were more difficult (7 and 8).

The logical sequence test presented a mean infit value of 0.97 (*SD* = 0.45), with a minimum value of 0.65 and maximum value of 2.22. Based on the results presented in Table 1, it was observed that the items were a good fit overall; only item 2 presented a value above the acceptable range. Mean outfit was 0.95 (*SD* = 0.63), ranging from 0.43 to 2.45. The outfit values for items 2 and 3 were respectively above and below the acceptable range (Table 1). Correlation values for LS ranged from 0.41 to 0.76, indicating adequate variability (Wright & Stone, 2004).

The infit and outfit results for item 2 indicate that this item should undergo a qualitative analysis, aiming to identify the interference of distracters that may have influenced the response pattern, given that both the difficulty level of the item and its correlation values were found to be within the expected standards. Another hypothesis, supported under the perspective of Gahyva and Hage (2010), is that the procedure has provided more a measure of cognitive processes (e.g. perception, memory, attention and meta-language) than of comprehension of the sequence of scenes shown in the test.

Also with respect to Table 1, results for the Logical-Verbal Organization test (LVO) indicated an item difficulty range of -1.09 to 0.81, considered easy. Three easier items were identified (items 2, 5 and 6), five of midrange difficulty (1, 3, 4, 7 and 8), and two with difficulty values higher than 0.50 (9 and 10). The mean infit value for this test was 1.02 (*SD* = 0.35), with indices ranging from 0.68 to 1.89, within the acceptable critical range. LVO outfit values (Table 1) ranged from 0.67 to 1.85; only one item (item 1) was shown to be outside the acceptable range. Correlation indices for this test

Table 1
Statistical Indices of Item Fit, Difficulty (B), Infits and Outfits, and Correlations for Items in MC, LS and LVO Tests

Test	Items	Item fit	*b*	Infit	Outfit	Correlation
MC	1	0.21	-0.22	1.33	1.5	0.34
	2	0.15	0.62	1.17	0.95	0.56
	3	0.16	0.53	1.1	1.07	0.53
	4	0.24	-0.48	1.11	0.84	0.4
	5	0.16	0.53	1.05	1.12	0.51
	6	0.17	0.25	0.77	0.68	0.59
	7	0.26	-0.61	0.78	1.22	0.38
	8	0.19	0.06	1.11	0.98	0.49
	9	0.28	-0.76	0.92	0.33	0.44
	10	0.19	0.09	0.77	0.61	0.57
LS	1	0.06	-0.48	0.77	0.56	0.56
	2	0.08	0.08	2.22	2.45	0.72
	3	0.07	-0.47	0.65	0.43	0.41
	4	0.08	-0.48	0.71	0.52	0.74
	5	0.08	0.09	0.81	0.68	0.73
	6	0.07	-0.05	0.93	0.73	0.74
	7	0.07	0.76	1.29	1.86	0.71s
	8	0.07	0.37	0.84	1	0.69
	9	0.07	0.14	0.82	0.75	0.73
	10	0.07	0.05	0.7	0.55	0.76
LVO	1	0.07	0.02	1.89	1.85	0.53
	2	0.1	-1.09	0.94	0.67	0.62
	3	0.08	-0.16	1.43	1.32	0.61
	4	0.07	0.06	0.89	0.8	0.72
	5	0.08	-0.26	0.79	0.69	0.71
	6	0.08	-0.22	0.77	0.72	0.72
	7	0.07	0.13	0.94	0.94	0.69
	8	0.07	0.06	0.68	0.75	0.74
	9	0.07	0.64	1.00	0.86	0.71
	10	0.07	0.81	0.85	0.78	0.74

Source: own work

ranged from 0.53 to 0.74, indicating variability in the data (Wright & Stone, 2004).

In the Story Interpretation (SI) test, there was variability in the difficulty values of items, which ranged from -1.67 to 0.99. Item grouping revealed that three were easier (3, 4 and 10), three were of midrange difficulty (1, 2 and 5) and four items were more difficult than the others (6, 7, 8 and 9).

Mean infit was 1.05 (*SD* = 0.16), ranging from 0.89 to 1.48. The mean outfit value was 0.88 (*SD* = 0.28), ranging from 0.58 to 1.41. Both indices are within the acceptable critical value range, showing that the test items presented good fit. Correlation values for this test ranged from 0.29 to 0.75, indicating variability (Wright & Stone, 2004).

The Completing Sentences test (CSe) presented difficulty values of -1.47 to 2.51. Results in Table 2 show that some items were easier (2, 4, 7 and 10), with difficulty values ranging from -1.47 to -0.69; some were midrange, namely 5 (-0.09), 6 (0.13) and 8 (-0.36); and some were harder (1, 3 and 9), with values ranging from 0.74 to 2.51. Mean infit was 1.11 (*SD* = 0.29), ranging from 0.76 to 1.15. Outfit ranged from 0.61 to 1.15, with a mean value

Table 2

Statistical Indices of Item Fit, Difficulty (B), Infits and Outfits for Items of SI, Cse and Cst Tests

Test	Items	Item fit	B	Infit	Outfit	Correlation
SI	1	0.12	0.27	1.48	1.41	0.5
	2	0.14	-0.16	0.97	0.7	0.55
	3	0.19	-0.80	1.03	0.63	0.43
	4	0.16	-0.48	1.14	0.58	0.48
	5	0.14	-0.10	1.02	1.25	0.51
	6	0.11	0.69	0.98	1.17	0.63
	7	0.11	0.99	0.94	0.77	0.75
	8	0.12	0.34	0.95	0.79	0.63
	9	0.11	0.92	0.89	0.92	0.71
	10	0.29	-1.67	1.1	0.58	0.29
CSe	1	0.14	0.91	1.28	1.15	0.52
	2	0.28	-0.98	0.81	0.74	0.39
	3	0.15	0.74	1.71	1.13	0.5
	4	0.35	-1.47	0.86	0.61	0.34
	5	0.2	-0.09	0.76	0.64	0.53
	6	0.19	0.13	1.13	1.06	0.46
	7	0.25	-0.69	1.32	0.84	0.38
	8	0.22	-0.36	1.31	0.89	0.41
	9	0.1	2.51	0.83	1.15	0.73
	10	0.25	-0.69	1.07	0.72	0.42
CSt	1	0.16	1.2	0.82	0.8	0.58
	2	0.16	1.14	0.9	0.85	0.55
	3	0.17	1.03	1.34	1.06	0.48
	4	0.19	0.68	1.02	0.99	0.43
	5	0.21	0.47	0.91	1.06	0.41
	6	0.28	-0.06	1.04	0.85	0.36
	7	0.29	-0.13	1.47	1.01	0.33
	8	0.49	-1.18	1.41	0.96	0.19
	9	0.57	-1.38	1.27	0.19	0.3
	10	0.7	-1.77	1.78	0.28	0.21

Source: own work

of 0.89 (*SD* = 0.21). The fit indices are within the acceptable critical ranges. According to Wright and Stone (2004), the correlation between difficulty of items and responses given by the subjects showed adequate variability.

The difficulty level of the Completing Stories test (CSt) showed values ranging from -1.77 to 1.2, and items could be grouped into easy ones (2, 8 and 9), with indices from -1.77 to -1.18; midrange-difficulty items (1, 3, 6 and 10) with values ranging from -0.13 to 0.68; and difficult items (4, 5 and 7), with difficulty values ranging from 1.03 to 1.20. Infit ranged from 0.82 to 1.78, with a mean value of 1.2 (*SD* = 0.29). Only one item (8) presented a value above the acceptable range (Table 2). The range of outfit values was from 0.19 to 1.06, with 0.81 as the mean (*SD* = 0.3). Two items, namely 8 and 9, presented outfit values below the acceptable critical range (Table 2).

It should be noted that the test items could be grouped into difficulty levels in all of the BILO*v3* tests, revealing a need to rearrange the items in version 4 of the battery, given that this is the objective of analyses of this nature, according to

Urbina (2007). For purposes of standardization, items sequenced by increasing difficulty may also support the establishment of a cut-off point.

A general analysis of the battery indicated that the CSe test had the items with the highest levels of difficulty, whereas the CSt test had the easiest items, although difficult items were detected in the MC, LS, LVO and SI tests. A possible hypothesis is that the difficulty of the CSe items is related to the fact that these items are not contextualized within a specific theme, and they also have one word omitted because they are organized under the Cloze Oriented System (Joly, 2007). This makes oral comprehension more difficult because language depends on the analysis of contingencies within a context of interaction (Araújo et al., 2010; Wagner et al., 2010). Looking specifically at receptive language, according to Johnson et al. (2009), it is related to semantic aspects because it enables one to infer meanings of the linguistic code based on contexts in which they occur.

Regarding the fact that some CSt items were the easiest in the battery, it should be noted that, on one hand, the presence of multimedia resources might have favored a better performance by test respondents. The reason for this is that the computer-based test makes it possible to assess complex cognitive processes and abilities used in problem solving (Bennett, 2001; Joly & Reppold, 2010; Wall, 2000). On the other hand, according to Crespo-Allende and Alfaro-Faccio (2010), Geytenbeek et al. (2010), Hincapié-Henao et al. (2008), subjects revealed adequate vocabulary, grammatical structure, and consequently adequate development of language, which favors a better performance in oral comprehension.

According to Linacre (2002), it should be considered that infit and outfit values above the standard range (> 1.5) are more critical than those below it (< 0.5) as regards the precision estimate that they provide. Lower than expected outfit values were observed in items of the MC (9), LS (13), LVO (1) and CSt (10) tests. These items should remain in tests and be submitted to qualitative studies in the future.

Precision estimates were calculated using the Kuder-Richardson coefficient, which is provided by the analysis performed by the Winsteps program (version 3.70.0, Linacre, 2012) on each test of the BILO*v3* battery. Indices ranged from 0.8 to 0.97, revealing good consistency (Table 3).

These values were higher than those reported in studies performed using different versions of BILO (Freitas, 2010; Istome, 2010; Joly, 2010; Joly & Dias, 2009; Joly & Piovezan, 2012; Joly et al., 2008; Joly et al., 2009; Joly, Reppold, & Dias, 2010; Soares et al., 2010). Therefore, according to the described analysis criteria, the observed presence of a single latent trait to be measured, and the described results, the items of all tests in the BILO*v3* battery showed good fit from a structural perspective, which provides evidence of internal construct validity (Embretson & Reise, 2000; Evers, Sijtsma, Lucassem, & Meijer, 2010; Tristán & Vidal, 2007; Van der Linden & Hamblenton, 2010).

Table 3
Internal Consistency as Measured by Kuder-Richardson Coefficient

Tests	Coefficient
Morpho-syntactic Comprehension	0.8
Logical Sequence	0.96
Logical-Verbal Organization	0.97
Story Interpretation	0.96
Completing Sentences	0.95
Completing Stories	0.82

Source: own work

Final Considerations

In light of the scarcity of instruments available for language assessment using psychometric characteristics, both internationally (Bishop et al., 2006) and in Brazil (Macedo et al., 2007), BILO is an important contribution for diagnosing receptive language disorders and the progress of linguistic pattern development in school-age children. It should be highlighted that BILO is one of the few computer-based batteries being studied in Brazil (Gurgel, 2010; Joly & Reppold, 2010).

We suggest that future studies be performed using the battery, so as to contemplate regional Brazilian samples with a view to standardizing the procedure. This could be based on standards targeted at typical development or for the measurement of diagnostic and prognostic criteria established by the DSM IVTR (American Psychiatric Association, 2002) for the assessment of language disorders, especially those related to oral language comprehension.

References

Almeida, A. R., & Joly, M. C. R. A. (2008). Estudo correlacional entre a Bateria Informatizada de Linguagem Oral (BILO) e Peabody. In L. Almeida, C. Machado, M. Gonçalves & A. P. P. Noronha (Eds.), *Avaliação psicológica: formas e contextos* (pp. 1-13). Braga: Psiquilibrios.

American Psychiatric Association. (2002). *Manual diagnóstico e estatístico de transtornos mentais - DSM-IVTR* (4th ed.). Porto Alegre: Artmed.

Araújo, M.V.M., Marteleto, M. R. F., & Schoen-Ferreira, T. H. (2010). Avaliação do vocabulário receptivo de crianças pré-escolares. *Estudos de Psicologia, 27*(2), 169-176.

Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives,* 9(5), 1-23. Available at http://epaa.asu.edu/epaa/v9n5.html

Bermúdez-Jaimes, M. E., & Sastre-Gómez, L. V. (2010). Falsa creencia y desarrollo semántico del lenguaje en niños de 2 a 4 años. *Universitas Psychologica,* 9(3), 849-861.

Bishop, D. V. M., Adams, C. V., & Rosen, S. (2006). Resistance of grammatical impairment to computerized comprehension training in children with specific and non- specific language impairments. *International Journal of Language & Communication Disorders, 41*(1), 19-40.

Crespo-Allende, N., & Alfaro-Faccio, P. (2010). Desarrollo tardío del language: la conciencia metapragmática en la edad escolar. *Universitas Psychologica,* 9(1), 229-240.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah, NJ: LEA.

Evers, A., Sijstma, K., Lucassem, W., & Meijer, R. R. (2010). The Dutch Review Process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10*(4), 295-317.

Ferreira, D. R. S. A., Ferreira, W. A., & Oliveira, M. S. (2010). Pensamento e linguagem em crianças com Síndrome de Down: Um estudo de caso da concepção das professoras. *Ciências & Cognição, 15*(2), 216-227.

Freitas, J. C. P. (2011). *Estudo correlacional da Bateria Informatizada de Linguagem Oral (BILOV1v3) com Teste Token* (Dissertação de Mestrado, Universidade São Francisco). Retrieved from http://www.usf.edu.br/mestrado/psicologia.html

Gahyva, D. L. C., & Hage, S. R. V. (2010). Intervenção fonológica em crianças com Distúrbio Específico de Linguagem com base em um modelo psicolinguístico. *Revista do Cefac, 12*(1), 152-160.

Geytenbeek, J., Harlaar, L., Stam, M., Ket, H., Becher, J. G., Oostrom, K., & Vermeulen, J. (2010). Utility of language comprehension tests for unintelligible or non-speaking children with cerebral palsy: A systematic review. *Developmental Medicine & Child Neurology, 52*(12), 1098. http://dx.doi.org/10.1111/j.1469-8749.2010.03833.x

Gurgel, L. G., Plentz, R. D. M., Joly, M. C. R. A., & Reppold, C. T. (2010). Instrumentos de avaliação da compreensão de linguagem oral em crianças e adolescentes: uma revisão sistemática da literatura. *Revista Neuropsicologia Latinoamericana, 2*(1), 1-10.

Hincapié-Henao, L., Giraldo-Prieto, M., Lopera-Restrepo, F., Pineda-Salazar, D. A., Castro-Rebolledo, R., Lopera-Vásquez, J. P., ... Lopera-Echeverri, E.

(2008). Transtorno específico del desarrollo del lenguaje en una población infantil colombiana. *Universitas Psychologica, 7*(2), 557-569.

Issa, G. M. P. (2008). *Estudos de Evidencias de Validade da Bateria Informatizada da Linguagem Oral – BILO* (Dissertação de Mestrado, Universidade São Francisco). Retrieved from http://www.usf.edu.br/mestrado/psicologia.html

Istome, A. C. (2010). *Bateria Informatizada de Linguagem Oral (versão 3): Características psicométricas para educação infantil e ensino fundamental.* Unpublished manuscript, Psicologia, Universidade São Francisco, Itatiba-SP, Brazil.

Istome, A. C., & Joly, M. C. R. A. (2010). Estudo correlacional do Teste Dinâmico de Leitura com o teste Wisc III. *Psicologia: Teoria e Prática, 12*(1), 43-58.

Johnson, K. N., Karrass, J., Conture, E. G., & Walden, T. (2009). Influence of stuttering variation on talker group classification in preschool children: Preliminary findings. *Journal of Communication Disorders, 42*(3), 195-210.

Joly, M. C. R. A. (2007). The validity of Cloze Oriented System (COS): A correlation study with an electronic comprehension test and a reading attitude survey [Special issue]. *Psicologia Escolar e Educacional, 11,* 49-59.

Joly, M. C. R. A. (2008). Bateria Informatizada de Linguagem Oral – BILO [Computer software]. São Paulo, SP: Núcleo de Avaliação Psicológica Informatizada.

Joly, M. C. R. A. (2009a). Bateria Informatizada de Linguagem Oral – BILOv3 (Version 3.0) [Computer software]. São Paulo, SP: Núcleo de Avaliação Psicológica Informatizada.

Joly, M. C. R. A. (2009b). Estudos com o sistema orientado de cloze para o ensino fundamental. In A. A. A. dos Santos, E. Boruchovitch & K. L. de Oliveira (Orgs.), *Cloze: um instrumento de diagnóstico e intervenção* (pp.103-142). São Paulo, SP: Casa do Psicólogo.

Joly, M. C. R. A. (2010). *Bateria Informatizada de Linguagem Oral (BILO)* (Technical report). São Francisco, Itatiba: Núcleo de Avaliação Psicológica Informatizada/ Universidade São Francisco.

Joly, M. C. R. A., & Dias, A. S. (2009). Evidências de validade de uma prova informatizada de Linguagem Oral – BILO. *Psicologia: Teoria e Prática, 11*(2), 50-68.

Joly, M. C. R. A., Martins, R. X., Souza, A. C. Z., Istome, A. C., & Santos, C. R. O. A. (2008). Bateria Informatizada de Linguagem Oral. In L. Almeida, C. Machado, M. Gonçalves & A. P. P. Noronha (Orgs.), *Avaliação Psicológica: formas e contextos* (pp. 121-140). Braga: Psiquilibrios.

Joly, M. C. R. A., & Piovezan, N. M. (2012). Estudo de validade correlacional e de critério da Bateria Informatizada de Linguagem Oral (BILO) com prova de raciocínio. *Estudos de Psicologia, 29*(4), 499-508.

Joly, M. C. R. A., Piovezan, N. M., Soares, C. A., Lopes, R. de M. M., & Martins, D. F. (2009, September). *Avaliação das características psicométricas da Bateria Informatizada de Linguagem Oral – BILOv2*. Poster presented at the III Congreso Latinoamericano de Psicología, Mexico D. F., Mexico.

Joly, M. C. R. A., & Reppold, C. T. (2010). *Estudo de testes informatizados para avaliação psicológica.* São Paulo, SP: Casa do Psicólogo.

Joly, M. C. R. A., Reppold, C. T., & Dias, A. S. (2010). Avaliação da linguagem oral de crianças paulistas e gaúchas pela Bateria Informatizada de Linguagem Oral (BILOv2). In C. Hutz (Org.). *Avanços em avaliação psicológica e neuropsicológica de crianças e adolescentes* (pp.175-206). São Paulo, SP: Casa do Psicólogo.

Joly, M. C. R. A., Welter, G. M. R., Martins, R. X., Silva, J. M., Montiel, J. M., Lopes, F., & Carvalho, M. R. (2005). Sistema de Avaliação para Testes Informatizados (SAPI): estudo preliminar. *Psic (São Paulo),* 6(2), 51-60.

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

Linacre, J. M. (2012). Winsteps (Version 3.70.0) [Computer software]. Retrieved from http://www.winsteps.com

Macedo, E. C., Firmo, L. S., Duduchi, M., & Capovilla, F. C. (2007). Avaliando linguagem receptiva via teste token: Versão tradicional *versus* computadorizada. *Avaliação Psicológica,* 6(1), 61-68.

Pasquali, L. (Org.). (1999). *Instrumentos psicológicos: manual prático de elaboração*. Brasília, DF: LabPAM / IBAPP.

Pasquali, L. (2007). *Teoria de resposta ao item: teoria, procedimentos e aplicações*. Brasília, DF: LabPAM/UnB.

Prieto, G. (1999). Procedimientos de construccion y analisis de tests psicometricos. In S. M. Wechsler & R. S. L. Guzzo (Orgs.), *Avaliação psicológica: perspectiva internacional* (pp. 57-100). São Paulo, SP: Casa do Psicólogo.

Rodríguez, V. A., Santana, A. M., & Caballero, A. A. (2012). Implicaciones clínicas del diagnóstico diferencial temprano entre Retraso de Lenguaje (RL) y Transtorno Específico del Lenguaje (TEL). *Universitas Psychologica, 11*(1), 279-291.

Smeekens, S., Riksen-Walraven, J. M., & van Bakel, H. J. A. (2008). Profiles of competence and adaptation in preschoolers as related to the quality of prent-child interaction. *Journal of Research in Personality, 42*(6), 1490-1499.

Soares, T. M., Fernandes, N. S., Ferraz, M. S. B., & Riani, J. L. R. (2010). A expectativa do professor e o desempenho dos alunos. *Psicologia: Teoria e Pesquisa, 26*(1), 157-170.

Sternberg, R. J. (2010). *Psicologia cognitiva* (M. R. Borges Osório, Trad., 5a. ed.). Porto Alegre, RS: Artmed.

Tremblay, P., & Small, S. L. (2011). From language comprehension to action understanding and back again. *Cerebral Cortex, 21*(5), 1166-1177.

Tristán, A., & Vidal, R. (2007, April). *Linear model to assess the scale's validity of a test*. [Online Submission]. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, USA.

Urbina, S. (2007). *Fundamentos da testagem psicológica.* Porto Alegre, RS: Artmed.

Van der Linden, W. J., & Hamblenton, R. K. (2010). *Handbook of modern Item Response Theory*. New York: Springer-Verlag.

Wagner, L., Gillespie, R., & Greene-Havas, M. (2010). Development in children's comprehension of linguistic register. *Child Development, 81*(6), 1678-1686.

Wall, J. E. (2000). Technology-delivered assessment: Diamonds or rocks? Greensboro, NC: ERIC Counseling and Student Services Clearinghouse. (ERIC ED 446 325).

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials.* Wilmington, DE: Wide Range, Inc.