Abstract
In this paper the Battery of Reading Assessment [Bateria de Avaliação da Leitura - BAL] is described and the results of the validity studies are presented. The BAL includes five tests organized into three areas of assessment: word reading; reading comprehension of texts (narrative and expository); and listening comprehension of texts (narrative and expository). Each test consists of vertically scaled forms for Portuguese elementary students from grades 1 to 4. Criterion-related validity evidence was gathered for the five tests by studying, through correlation and regression analyses, the association between each test score and the results on other tests assessing reading, memory and vocabulary and on teachers’ assessment. The dimensionality of each test was studied through confirmatory factor analyses. The results provided evidence of statistically significant correlations of the tests with the external criteria and evidence of an acceptable fit for the one-factor solution for all tests.

Resumo
Bateria de Avaliação da Leitura: Estudos de validade. Nesta comunicação são apresentados os estudos de validade da Bateria de Avaliação da Leitura (BAL) (PTDC/PSI-EDU/098592/2008). A BAL inclui cinco testes organizados em três domínios de avaliação: (a) a leitura de palavras; (b) a compreensão de textos (narrativos e informativos) na modalidade de leitura; (c) a compreensão de textos (narrativos e informativos) na modalidade oral. Cada teste é composto por versões equalizadas verticalmente para alunos do 1.º ao 4.º anos de escolaridade. Foi recolhida evidência de validade de critério para os cinco testes estudando, através de análises de correlação e regressão, a associação entre os resultados de cada teste e os resultados em outros testes de leitura, de memória e de vocabulário e, ainda, na avaliação dos professores. Estudou-se a dimensionalidade dos testes com recurso à análise fatorial confirmatória. Os resultados evidenciaram correlações estatisticamente significativas entre os testes e os critérios externos e permitiram confirmar a unidimensionalidade dos testes.
Introduction

The results of international literacy studies like Reading Literacy (I.E.A., 2008), The Progress in International Reading Literacy Study (N.C.E.S., 2008) and P.I.S.A. 2006 (O.E.C.D., 2010) have made it clear the differences in reading literacy competence among students from diverse educational systems. These results have led some countries to initiate efforts to define national (or international, in the case of the Commonwealth) reference standards and criteria for reading comprehension performance. In Portugal, similar concerns have been expressed.

In 2006, the Portuguese Government initiated the National Plan for the Teaching of Portuguese in the Elementary Education (Ministério da Educação, 2006; Sim-Sim, 2007). One of the aims was to “improve the levels of reading comprehension, writing and oral expression in all elementary schools in a period from 4 to 8 years through the modification of language teaching practices”. Nevertheless, this implied that teachers were familiar with differentiated evaluation practices and with the existence of measures that allowed for the monitoring of learning and anchoring of pedagogical decisions. In 2007, Sim-Sim and Viana singled out the Portuguese reading evaluation tools for the first six school grades (including, but not limited to, the elementary levels). This report evidenced the lack of valid and reliable instruments to evaluate reading comprehension and it put forward a list of descriptors in reading proficiency for the first six school grades.

Some tests were developed, in previous studies, to assess word recognition or reading comprehension (Pereira, Ribeiro, Viana, & Maia, 2008; Vale & Bartelli, 2006; Viana, Pereira, & Teixeira, 2003; Viana & Ribeiro, 2006, 2007). However, this scattered set of tests does not permit the analysis of individual trajectories of reading development. The Test of Reading Comprehension (Cadime, Ribeiro, & Viana, 2012) is an exception since it includes vertically scaled test forms for students from grades 2 to 4 that allow the monitoring of the reading comprehension growth across these school levels. The test includes different types of texts and it presents good psychometric properties (Cadime et al., 2013), however, it does not allow to differentiate students’ reading comprehension performance according to different types of texts. This is an important aspect because research (e.g., PISA 2006) indicates that individual differences in reading performance are influenced by the text genre.

The Battery of Reading Assessment [BAL – Bateria de Avaliação da Leitura] (Ribeiro & Viana, 2014a) was developed in order to allow the assessment of core reading skills (word reading, listening and reading comprehension). The aim of this paper is to present an overall description of the BAL, including results concerning reliability and validity.
Description of the Battery of Reading Assessment

The BAL includes 18 tests that assess word reading (4 test forms), listening comprehension of narrative (4 test forms) and expository texts (4 test forms) and reading comprehension of narrative (3 test forms) and expository texts (3 test forms).

Item analysis was performed within the Item Response Theory framework, more specifically within the Rasch Model for dichotomous items. This model was also used to construct vertically scaled forms to assess students from grades one to four. For a detailed description of vertical scaling procedures, see Cadime, Ribeiro and Viana (2014). Table 1 presents a brief description of BAL.
<table>
<thead>
<tr>
<th>Tests</th>
<th>Test of Word Recognition</th>
<th>Test of Listening Comprehension</th>
<th>Test of Reading Comprehension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goals</td>
<td>Word reading assessment</td>
<td>Listening comprehension of narrative and expository texts assessment</td>
<td>Reading comprehension of narrative and expository texts assessment</td>
</tr>
<tr>
<td>Authors</td>
<td>Viana, Ribeiro, Vale, Chaves-Sousa, Santos, &amp; Cadime (2014 a, b, c, d, e)</td>
<td>Viana, Ribeiro, Santos, &amp; Cadime (2014 a, b, c, d, e, f, g, h, i, j)</td>
<td>Ribeiro, Viana, Cadime, Santos, &amp; Spinillo (2014 a, b, c, d, e, f, g, h)</td>
</tr>
<tr>
<td>Publisher</td>
<td>CEGOC-TEA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>Grades 1 to 4</td>
<td>Grades 2 to 4</td>
<td></td>
</tr>
<tr>
<td>Administration</td>
<td>Individual, untimed</td>
<td>Individual or collective, untimed</td>
<td></td>
</tr>
<tr>
<td>Format</td>
<td>Digital</td>
<td>Paper and pencil</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td>Items</td>
<td>30 items (isolated words)</td>
<td>30 multiple choice items per test and per version</td>
<td>TCTML-n: 33 multiple choice items per version TCTML-i: 27 multiple choice items per version</td>
</tr>
<tr>
<td>Items characteristics</td>
<td>Words were selected according to the following criteria: extension, frequency, regularity, syllabic structure</td>
<td>The following reading levels are assessed: reorganization, literal, inferential and critical comprehension</td>
<td></td>
</tr>
<tr>
<td>Norms</td>
<td></td>
<td>Percentiles by grade</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cut-off points</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Test of Word Recognition [TLP - Teste de Leitura de Palavras]**

TLP is a test that measures single words’ oral reading. The words were chosen considering length, frequency, regularity and syllabic structure. For frequency estimation ESCOLEX database was used (Soares et al., 2014). Four vertically scaled forms were built to assess students from grades one to four. Each test form includes 30 items (single words) that are displayed through a computer application. Test words from each form are presented in isolation in a randomized order in a computer screen. The test administration is individual and untimed. During the test application, the word reading accuracy (correct/incorrect) is registered by the evaluator. At the end of the task, the computer application exports the assessment results to an Excel file (Viana, Ribeiro, Vale, Chaves-Sousa, Santos, & Cadime, 2014a).

**Tests of listening and reading comprehension with narrative and expository texts**

Measures of reading comprehension have been classified as online or offline. With online measures, the information is collected during reading, while with offline measures, the information is collected after reading (Lencastre, 2003).

The two tests of Listening Comprehension (Viana, Ribeiro, Santos, & Cadime, 2014, a, f) and Reading Comprehension (Ribeiro, Viana, Cadime, Santos, & Spinillo, 2014, a, e) are offline measures. The reading comprehension tests require the children to read silently the text and the items, while in the listening comprehension tests children listen to the text and the items.

**Texts**

In the construction of the comprehension assessment tools, the option was to create different tests for narrative and expository texts, both in the oral and the reading modality. Sim-Sim and Viana (2007) guidelines were adopted regarding the structure and text length.

All texts were original and authored by Portuguese children’s literature writers (Ana Maria Magalhães, Isabel Minhós Martins and Carla Maia de Almeida) and by a scientific divulgation writer (Jorge Gomes) ensuring that the texts’ vocabulary and length fitted children’s grade level skills. The option of using original and unpublished texts was envisioned to exclude the possibility of an effect of previous knowledge of some students, which would place them in advantage relatively to their reference group. Four texts were included in each test form of TCTML-n and TCTML-i. Their extension increases as the grade levels progresses. Each text incorporated the previous grade text as well as new content: 4th grade texts included the 3rd grade texts and a new part; the 3rd grade texts comprised the 2nd grade texts and, again, a new extension; and finally, the 2nd grade texts contained the 1st grade texts and were also expanded. This conception enabled the students’ assessment
using a text tailored to their ages or grade levels and it reduced the possibility of obtaining a ceiling effect in the higher grades. Moreover, this format allowed building anchor items, i.e., common items between the test forms for the adjacent grades. Anchor items are required to perform the vertical scaling of the different test forms' scores.

Format of the items and comprehension levels

In the comprehension tests, the items were multiple-choice questions with three options, with one correct. Using three options has been shown to be as effective as formats using four or five options in the discrimination of performance levels, it does not affect the measures’ reliability and it is time-saving in administration (Delgado & Prieto, 1998; Rodriguez, 2005). In addition to psychometric issues, more than three alternatives could overload the working memory needed for handling the tests of Listening Comprehension. For these reasons, and in order to keep the equivalence of the two assessment modalities (reading versus listening), the choice was made to have three answer alternatives in all the items that compose the four comprehension tests. Both the items and the corresponding alternative answers were reviewed by linguistic experts. The order of presentation of the correct alternative in each item was randomised.

Items assess reorganization (the ability to synthesize or schematize the text content), literal comprehension (the ability to comprehend explicit information in the text), inferential comprehension (the ability to draw conclusions about information that is implicitly stated in the text) and critical comprehension (the ability to make judgments or give opinions about the text).

Silent reading versus listening

TCTML-n and TCTML-i tests are presented in a pencil-and-paper format; TCTMO-n and TCTMO-i tests are presented in a digital format. In the reading comprehension tests, the student must read silently both the text and the questions and mark the chosen option in the answer sheet. These two tests can be administered individually or collectively. In the listening comprehension tests, the student listens to the texts and the items that are presented orally. The tests' administration is individual or collective. When the administration is performed collectively (in group), each child must have a computer and use headphones. At the end of the task, the computer application exports the assessment results to an Excel file.
Reliability and validity

Reliability was examined by computing Rasch coefficients for the Person Separation Reliability (PSR), Item Separation Reliability (ISR), and the Kuder-Richardson formula 20 (KR20). All three coefficients are expressed on a scale ranging from 0 to 1. High reliability coefficients indicate low levels of measurement error; therefore, values closest to 1 are desirable (Bond & Fox, 2007; de Ayala, 2009). Evidence regarding the criterion-related validity was analyzed by correlating the results with the other tests of the BAL and with the results in tests of reading comprehension (Cadime, Ribeiro, & Viana, 2012), word recognition (Viana & Ribeiro, 2010), memory (Wechsler, 2003), vocabulary (Wechsler, 2003) and fluency (Carvalho, 2010). They were also correlated with the teachers' evaluation of reading skills (decoding, listening and reading comprehension).

The study of the reliability (ISR, PSR and KR20) and the factorial structure of each test form, with confirmatory factor analysis, was performed with students of grades 1 to 4. The PSR, the ISR and the KR20 were higher than .74. Confirmatory factor analysis indicated a one-factor structure for each test form. Correlations were moderate with the external criteria. No statistical significant differences were found between boys and girls in the tests that compose the BAL.

The use of the BAL in research, clinical and educational contexts

The development of the BAL can help overcome the Portuguese needs concerning reading assessment, with implications in the psychological practice and in reading research – which has a significant importance since the internationalization of the investigation requires studies in which valid instruments are used. The BAL includes a "set of tests that allow a broad and rigorous assessment of reading" (Ribeiro & Viana, 2014b, p. 12).

According to Ribeiro and Viana (2014b) the use of the tests of the BAL will allow:

to determine the discrepancies between individual performance in word reading, listening and reading comprehension with the proper normative group. It will be possible to test hypothesis regarding the influence of word reading and listening comprehension on students' performance in reading comprehension and its relation with learning difficulties. The forms of the five tests of the BAL were subject to a procedure of vertical scaling. This option permits the evaluation of intra-individual changes over time. There is, therefore, the possibility of monitoring gains in time and the approximation to the expected standards for each school grade. This is particularly important to assess the short-term and mid-term effects of intervention programs in students with difficulties in word reading, reading comprehension and listening comprehension (pp 12-13).
References


Cadime, I., Ribeiro, I., & Viana, F. L. (2012). Teste de Compreensão da Leitura (TCL) (Teste Psicológico) [Reading comprehension test (psychological test)]. Lisboa: CEGOC-TEA.


TLP-3: Teste de Leitura de Palavras 3.º ano [Test of Word Recognition-3]. Lisboa: 
CEGOC-TEA.

TLP-4: Teste de Leitura de Palavras 4.º ano [Test of Word Recognition-4]. Lisboa: 
CEGOC-TEA.

edição (WISC – III) [Wechsler Intelligence Scale for Children-III]. Portuguese 

This research was supported by Grant COMPETE: FCOMP-01-0124-FEDER-010733 from FCT 
(Fundação para a Ciência e Tecnologia) and the European Regional Development Fund (FEDER) 
through the European program COMPETE (Operational Programme for Competitiveness Factors) 
under the National Strategic Reference Framework (QREN).