Santos et al. Psicologia: Reflexão e Crítica (2016) 29:38
DOI 10.1186/s41155-016-0044-6

Psicologia: Reflexão e Crítica

RESEARCH Open Access

An application of the Rasch model to reading comprehension measurement

Sandra Santos, Irene Cadime, Fernanda Leopoldina Viana, Gerardo Prieto, Séli Chaves-Sousa, Alina Galvão Spinillo and Iolanda Ribeiro*

Abstract

An effective reading comprehension measurement demands robust psychometric tools that allow teachers and researchers to evaluate the educational practices and track changes in students' performance. In this study, we illustrate how Rasch model can be used to attend such demands and improve reading comprehension measurement. We discuss the construction of two reading comprehension tests: TRC-n, with narrative texts, and TRC-e, with expository texts. Three vertically scaled forms were generated for each test (TRC-n-2, TRC-n-3, TRC-n-4; TRC-e-2, TRC-e-3 and TRC-e-4), each meant to assess Portuguese students in second, third and fourth grade of elementary school. The tests were constructed according to a nonequivalent groups with anchor test design and data were analyzed using the Rasch model. The results provided evidence for good psychometric qualities for each test form, including unidimensionality and local independence and adequate reliability. A critical view of this study and future researches are discussed.

Keywords: Rasch model, Vertical scaling, Reading comprehension, Assessment

Background

When developing robust reading comprehension tests, researchers must consider the specific challenges concerning the assessment of a skill that is expected to develop over the course of schooling (Taylor et al. 2005; Taylor & Pearson, 2005) and that is directly related to the development of other skills, including listening comprehension, word reading, vocabulary knowledge and general cognitive abilities (Cain, 2010). For those reasons, to improve reading assessment quality, it is recommended that reading assessment should allow for the monitoring of the students' progress across grade levels and the diagnosis of reading difficulties (Cain, 2010).

Several reading comprehension tests have been constructed to assess comprehension as a function of age or academic grade, so it is possible to compare a student's performance with his/her normative group (e.g., Stanford Diagnostic Reading Test, 4th Ed., Karlsen & Gardner, 1996; Woodcock et al. 2004). However, when students take different tests across academic grades, intra-individual differences in developmental profiles cannot be compared because test performance estimates are on different metrics and there is no functional relationship between scales. Using the same test across a wide range of grades raises a problem; depending on test difficulty, the results can be affected by severe ceiling effects in higher grades or extreme floor effects in lower grades. To address challenges associated with constructing tests that are sensitive to intra-individual changes, items from different tests must be placed in a single metric. This can be achieved using vertical scaling, i.e., the construction process of a scale meant to allow growth in a determined ability across developmental phases to be measured (de Ayala, 2009).

This paper presents two original tests that address the aforementioned constraints: the Test of Reading Comprehension of Narrative Texts (TRC-n) and the Test of Reading Comprehension of Expository Texts (TRC-e). Three vertically scaled forms of each test were developed to assess reading comprehension of elementary Portuguese students, respectively, in second, third and fourth grade of elementary school. The TRC-n is composed of test forms TRC-n-2, TRC-n-3 and TRC-n-4 and the TRC-e comprises the test forms TRC-e-2, TRC-e-3 and TRC-e-4. These two reading comprehension tests were constructed taking into account the recommendations of

* Correspondence: iolanda@psi.uminho.pt
Escola de Psicologia, Universidade do Minho, Braga, Portugal

several authors (e.g., Afflerbach, 2004; Eason et al. 2012; Hess, 2007; Sweet, 2005) to allow comparisons of results across grades (intraindividual level), as well as within each grade (interindividual level).

Issues in reading comprehension measurement

Reading comprehension is defined as the process of simultaneously extracting and constructing meaning through the involvement with written text (RAND Reading Study Group, 2002). The reader activates multiple cognitive processes at the word, sentence and text levels to integrate the information from the text with his previous knowledge and, therefore, to construct a mental representation of the text message (Chen & Vellutino, 1997; Kintsch, 1998; Perfetti et al. 2005).

Research has shown that factors within the reader, factors associated with the text being read and the tasks presented affect the nature of reading comprehension. Reader's variables include linguistic knowledge, cognitive abilities, motivational aspects and world knowledge (cultural, social, affective). Text variables comprise text content, text type, text organization, and text readability. Reading tasks vary in their level of complexity according to the text, the test-tasks (e.g., multiple choice items, true/false questions), the reader, and the interaction among them (Alderson, 2000; RAND Reading Study Group, 2002). Also, the nature of the task presented to the reader must be considered since different tasks (e.g., text reproduction, question-answer, cloze task) access different facets of reading comprehension (Spinillo et al. 2016).

Regarding the text types, the literature indicates that differences in achievement can, to some extent, be traced to the use of narrative or expository text in reading comprehension tests (Eason et al., 2012; Mullis et al. 2012; RAND Reading Study Group, 2002). One way of dealing with this issue is to include different types of texts in the same test as it can be observed in some tests like the TCL, a Portuguese reading comprehension test (Cadime et al., 2013) and the ACL, a Spanish reading comprehension test (Català et al. 2001). Another alternative is to construct independent tests, one for each type of text, as in the Progress in International Reading Literacy Study (Mullis et al., 2012). According to some researchers (Hess, 2007; Sweet, 2005), this alternative allows to obtain differentiated results for narrative and expository texts. Different scores may contribute to characterize individual achievement in reading comprehension, and guide instruction to specific text types.

Whether the texts are narrative or expository, tests must assess different comprehension levels, which involve the abilities to: (a) identify literal information; (b) make inferences; (c) reorganize information and; (d) make critical judgments (Alderson, 2000). Literal comprehension (LC), inferential comprehension (IC), reorganization (R) and critical comprehension (CC) are terms used to differentiate these levels, respectively (Barrett, 1976; Català et al., 2001).

Reading comprehension assessment has focused mainly on literal and inferential comprehension (e.g. Andreassen & Braten, 2010; Best et al. 2008; Goff et al. 2005; Hess, 2007), whereas reorganization and critical comprehension levels are considered as sub-levels of inferential comprehension. As guideline for test item development, it is useful to consider separate levels understanding in order to gauge the spectrum of skills associated with reading comprehension (Eason et al., 2012).

Although different reading comprehension levels can be identified in different taxonomies, they all characterize the demands of the tasks where reading comprehension is the underlying latent factor (Basaraba et al. 2013; Ozuru et al. 2008). A study conducted by Cadime and colleagues (2013) assessed the construct validity of a reading comprehension test for Portuguese students in the second to fourth grades also validated this hypothesis. In this test the four levels of comprehension (LC, IC, R and CC) were considered and the results from the confirmatory factorial analysis revealed a one-factor structure for the reading comprehension construct in every school grade.

Thus, as guidelines, the construction of reading comprehension tests should include scores related to different text types (mainly narrative or expository), consider the different comprehension levels and allow analyzing inter- and intra-individual changes over time.

The Rasch model

Item response theory models have been widely used in measurement applications in social areas such as language and educational testing (Boone & Scantlebury, 2006; Boone et al. 2011; Wilson & Moore, 2011). One of the models associated to Item Response Theory is the Rasch model, according to which the probability of a person responding correctly to an item depends on the difficulty of the item and on the person's ability regarding the latent trait (Rasch, 1980).

In Rasch model analysis, two parameters are estimated: a difficulty parameter for each item (b_i) and an ability parameter for each person (θ). These are placed on a single logit scale and a continuum is constructed on which the items and persons are ordered according to their respective parameter values. Greater distances between the person value x and the item value y (in favor of the person) indicates greater chances of giving a correct response to the item. Greater distances in favor of the item (i.e., the value for the item difficulty is higher than the person ability) indicate lower probability of a correct response. Items that are too easy or too difficult for a particular person or group are less informative than

items that are located approximately at the same level of the person ability along the continuum (for details, see Bond & Fox, 2007).

The process of aligning metrics by placing all the items parameters on the same scale is known as linking (de Ayala, 2009). When the item parameters are used to estimate the persons' ability (θ), the person parameter estimations are also placed on the same metric. The process of measuring person estimations on the same scale is called equating. The construction process of a scale meant to allow growth in a determined ability across developmental phases to be measured is known as vertical scaling or vertical equating. Vertical scaling is only possible if the test forms support similar score inferences, assess the same construct for the same target population and are designed to be as similar as possible in content and statistical characteristics, except difficulty, which is expected to increase with academic grade levels (Kolen & Brennan, 2010).

This paper aimed to illustrate how Rasch model can contribute on reading comprehension measurement through multiple-choice tests. Three specific purposes were defined: (a) to examine the psychometric properties of the items; (b) to assess the test forms dimensionality, local independence and reliability, and; (c) to perform vertical scaling of the test forms of the TRC-n and the TRC-e.

Method

Participants

All participants were native speakers of European Portuguese attending primary schools located in urban and rural areas in Portugal. One sample of 702 students took part in the study of the TRC-n. They were divided into three groups: 230 s graders (52.2 % were male, 57.8 % attending urban schools), 239 third graders (50.6 % were male, 63.2 % attending urban schools) and 233 fourth graders (56.7 % were males, 52.8 % attending urban schools). Another sample of 742 students took part in the study of the TRC-e. They were also divided into three groups: 252 were second graders (46.6 % were male, 77.3 % attending rural schools), 222 third graders (50.7 % were males, 74.9 % attending rural schools) and 268 fourth graders (51.5 % were males, 76.1 % attending rural schools).

Materials

Each test form of the TRC-n comprised a booklet of texts and a worksheet with items. The booklet contained four narrative texts that were common to the three test forms, however, their length increased along with grade level: for the second graders the texts ranged from 138 to 289 words, for the third graders from 363 to 544 words and for the fourth graders the texts ranged from 495 to 915 words (Appendix B). All texts were original and authored by Portuguese writers of literature for children. The authors were asked to write texts that would be a complete and coherent text and also be part of the text presented in the following grade. Thus, the texts given to the second graders were part of the texts presented to the third graders, which, in turn, were both contained in the texts presented to the fourth graders. This model enabled the construction of common items that were necessary to perform vertical scaling of the test forms. This was also the underlying reason for the need of multiple texts. For instance, in the second grade, each text *per se* was not sufficient to build an adequate number of items to accurately assess comprehension; however, the summation of multiple texts enabled such an assessment (see Appendix A and Appendix B).

Items were multiple-choice questions with three options. They were developed to assess LC, IC, R or CC and underwent an analysis of the content by linguistics and reading comprehension experts. Each test form included unique items and items that were common between the test forms for the adjacent grades (see "Initial pool of items" column in Table 1).

Test material in the TRC-e was similar to the material in the TRC-n, except for the fact that the booklets contained four original expository texts. The length of these texts increased along with grade level: for the second graders the texts ranged from 173 to 196 words, for the third graders from 351 to 502 words and for the fourth graders the texts ranged from 701 to 940 words. Each test form contained an increasing number of items in subsequent grades, including unique and common items (see "Initial pool of items" column in Table 1).

The syntactic complexity in the TRC-n and TRC-e texts also increased along with grade level, taking into account grammatical structures: simple or complex sentences, coordination or subordination, word order, and anaphoric chains. The specifications of the texts' extension and complexity followed the recommendations of the report prepared by Sim-Sim and Viana (2007), supported by the National Reading Plan (Ministry of Education and Science), and the guidelines of the curricular benchmarks for Portuguese language, a reference document for teaching and learning and for internal and external evaluation (Buesco et al. 2015).

Design and procedures

A nonequivalent groups with anchor test design, also known as common-item nonequivalent groups design was used (de Ayala, 2009; Kolen & Brennan, 2014). This design involves having at least two groups that differ in ability level responding to different test forms. The groups are assumed to be nonequivalent because they consisted of students in different grade levels with

Table 1 Items in each test form of the TRC-n and TRC-e

Test form	Initial pool of items							Selected items for the final test forms						
	T	Items per level				U	A	T	Items per level				U	A
		LC	IC	CC	R				LC	IC	CC	R		
TRC-n-2	41	13	21	2	5	23	18	27	8	14	2	3	21	6
TRC-n-3	47	11	26	7	3	13	6[a],12[b],16[c]	27	6	15	4	2	16	5[a],1[b],5[c]
TRC-n-4	57	16	30	4	7	29	28	27	6	15	4	2	21	6
TRC-e-2	36	16	15	2	3	16	20	33	15	14	2	2	25	8
TRC-e-3	53	21	25	2	5	7	3[a],17[b], 26[c]	33	9	18	5	1	17	8[a], 8[c]
TRC-e-4	79	24	41	3	11	36	43	33	8	15	7	3	25	8

Note. T total number of items, *LC* literal comprehension, *IC* inferential comprehension, *CC* critical comprehension, *R* reorganization, *U* number of unique items, *A* number of anchor items

[a]The number of common items shared by the second and third grade test forms

[b]The number of common items shared by the second, third and fourth grade test forms

[c]The number of common items shared by the third and fourth grade test forms

different reading abilities. As mentioned, the test forms included a set of items that were common between forms for adjacent grades, and a set of items unique to each form. Common items were interspersed among the other items in each test form (Fig. 1). The values of the common items, also known as anchor items, are crucial to align the metrics from different test forms, in order to compare scores.

Students of the first sample completed the TRC-n-2, TRC-n-3 or TRC-n-4, and students of the second sample completed the TRC-e-2, TRC-e-3 or TRC-e-4, according to their grade. Tests were administered collectively, in the classroom, by trained psychologists. Legal authorization for data collection was solicited from the Portuguese Ministry of Education and school boards and informed consent for test administration was previously collected from students and their parents or legal guardians.

Data analyses

Data were analyzed using the Rasch measurement software Winsteps 3.72.0 (Linacre, 2011) and The Statistical Package for the Social Sciences 20.0 (IBM Corporation) was also used to compute statistics and evaluate between-grade differences in the scaled scores obtained on each form.

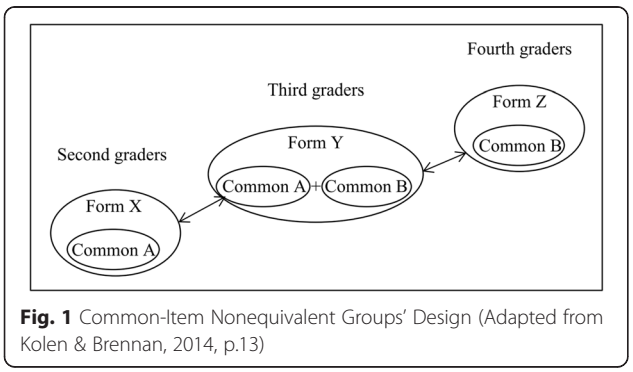

Fig. 1 Common-Item Nonequivalent Groups' Design (Adapted from Kolen & Brennan, 2014, p.13)

The psychometric properties of the items were analyzed considering the following criteria: (1) adequacy of the difficulty of the items to the group being assessed; (2) fit of the data to the model; (3) association between the score in one item and the latent trait assessed; and (4) adequacy of the options in each item. Person (θ) and item parameters (b_i) were computed and compared and a visual representation of these parameters was displayed in the form of person-item maps. This analysis allowed to detect items that were of inadequate difficulty (too easy or too difficult) for each grade for the latent trait assessment or if there was any region of the latent trait that was not measured by any item (Boone & Scantlebury, 2006). Two fit statistics were calculated for each person and each item to evaluate the degree of fit of the model to the data: (i) infit mean square statistic focuses on unexpected responses near a person or item measure, and (ii) outfit mean square statistic places more emphasis on unexpected responses far from a person or item measure, being more sensitive to outliers. Infit and outfit values should not be higher than 1.5 (Linacre, 2002). Point-measure correlations were computed for each item within each test form to evaluate the degree of association between the score in one item and the latent trait assessed by the test form. A mean ability value for persons who chose each option relative to each item was also calculated. For each item, the highest mean value is expected to be observed for the group of persons who select the correct answer option (Linacre, 2011).

Given that a nonequivalent groups with anchor test design was used, the item selection was performed at the same time that test forms were linked. The fixed item parameter method was used to link the test forms. This procedure calibrates each test form separately and sequentially. A test form is selected as the reference test and the measure values of the anchor items obtained by its calibration are used to perform the calibration of a

target test form (de Ayala, 2009). According to this procedure, the values of the common items are fixed to the values obtained by the calibration of the reference test and are not re-estimated in the target test calibration, thus placing the target test form on the measurement scale of the reference test form (Kolen & Brennan, 2010). Anchor items that are used in test scaling or linking should ideally be distributed along the difficulty range of the test, be representative of the test content and of reasonably stable difficulty across test forms (Dorans et al. 2011; Huynh & Meyer, 2010). To evaluate the stability of the common items' difficulty between adjacent grades, displacement measures should be estimated for each anchor item to check discrepancies between the anchor value and the parameter that would have been estimated for the target test form if this parameter was unanchored (Linacre, 2011). The values of the anchor items' displacement should be less than 0.50 logits, but they can assume values as large as 1.0 logits without causing much impact on measurement, as mentioned by Linacre (2011).

In this study, the reference test forms were the TRC-n-2 and the TRC-e-2 to link the second and third grade forms of each test and the reference tests to link the third and fourth grade forms of each test were the TRC-n-3 and the TRC-e-3. The TRC-n forms were linked according to the following steps: (a) calibration of the TRC-n-2; (b) elimination from the TRC-n-2 of the items with inappropriate psychometric characteristics and re-estimation of the parameters in a new calibration; (c) calibration of the TRC-n-3 fixing the parameters of the common items in the values obtained in the last calibration of the TRC-n-2; (d) elimination from the TRC-n-3 of the items with inappropriate psychometric characteristics and re-estimation of the parameters in a new calibration; (e) analysis and selection of the anchor items to link the TRC-n-2 and the TRC-n-3 and reduction of the number of unique items in each test form considering the values obtained in the last calibration of each one. Unique items were selected according to three criteria: (a) the spread of difficulty – items that were distributed spread along the continuum of ability of each grade sample were selected; (b) redundancy – the number of items of similar difficulty levels was reduced by discarding some redundant items; and (c) the comprehension level that was assessed – the proportion of items of each type that was present in the initial pool of items was maintained in the final pool of items. When two or more items were of similar difficulty levels and they measured the same comprehension level, the item with a higher point-measure correlation was retained.

The same steps were adopted in the calibration and anchor item selection of the TRC-n-4. The whole process was replicated to link the TRC-e test forms.

A final set of calibrations was performed, using the fixed item parameter method, so that the test forms were again linked. Only the unique and anchor items selected in the previous phases were considered in this calibration. One-way analyses of variance (ANOVAs) and post hoc comparisons tested the between-grade differences in the scaled scores obtained on each form.

Unidimensionality and local independence of the items of the TRC-n and TRC-e final forms were tested. Unidimensionality assumes that the performance of one person in the observed variables is dependent of one specific single latent variable of the person (de Ayala, 2009). Local independence indicates that the performance of one person on any item is determined solely by his or her level on the latent variable and does not depend on external variables (Bond & Fox, 2007). Unidimensionality of each TRC-n and TRC-e form was tested using principal component analysis (PCA) of the linearized Rasch residuals. This analysis enables the identification of clusters of residuals that share a large amount of common variance. These clusters are referred to as secondary dimensions, which should have at least the strength of two items to be considered a possible separate dimension from the dimension tapped by the other items. Therefore, eigenvalues less than 2.0 for secondary dimensions support the unidimensionality assumption.

Correlations between the items' linearized Rasch residuals within each test form were computed to examine the requisite of local independence of the items. Residuals that are highly correlated indicate that performance on an item does not depend only on the individuals' ability level (θ), but may be "contaminated" by the response to another item. Correlations higher than .70 may indicate that items are locally dependent (Linacre, 2011).

Reliability of each final form of the TRC-n and TRC-e was examined by computing Rasch coefficients for the Person Separation Reliability (PSR) and 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).

Results

The description of the results is organized by reading comprehension test.

Test of reading comprehension of narrative texts (TRC-n)

On the TRC-n-2 initial pool of items, item difficulty in logits (b_i) ranged from −2.25 to 1.63, and person ability ranged from −1.16 to 2.84. The minimum person ability value exceeded the minimum value of items difficulty, meaning that some items were excessively easy for all

grades. On the TRC-n-2, items RB2.7^{LC}, RB2.8^{LC}, RB2.12^{LC} and AN2.1^{LC} exhibited difficulty values lower than the minimum value for person ability (minimum = −1.16), meaning that they were very easy for second graders. Infit and outfit statistics for the items of the TRC-n-2 did not exceed the reference value of 1.5. Regarding the fit statistics for the person results in TRC-n-2, none of the students presented an infit value greater than 1.5 and only six second graders (2.6 % of the second grade sample) obtained outfit values greater than 1.5. Point-measure correlations in the TRC-n-2 ranged from -.03 for item LB2.7^{R} to .52 for item RB2.10^{LC}. Negative point-measure correlation coefficients were found for items RB3^{IC} and LB2.7^{R} on the TRC-n-2. For these items and items LB2.5^{R}, the highest mean ability value was not obtained by students who chose the correct answer option, suggesting that the students with greater reading comprehension abilities chose an incorrect alternative.

Items of inadequate difficulty, fit statistics greater than 1.5, negative point-measure correlations and/or problems in the answer options (seven items) were removed from TRC-n-2 and the test form with 34 selected items was recalibrated. Table 2 shows the psychometric properties of the items of the TRC-n-2. None of the items revealed inappropriate psychometric characteristics in this new analysis. Regarding the fit statistics for the person results in TRC-n-2, no changes were observed.

The TRC-n-3 initial pool of items was then linked to the TRC-n-2 (34 items). The calibration results show that item difficulty ranged from −1.31 to 2.67 and person ability ranged from −1.34 to 3.30. None of the items was identified as too easy or too difficult for the third grade. Infit statistics for the items of the TRC-n-3 did not exceed the reference value of 1.5 and only item AN12^{CC} had outfit values higher than 1.50. Regarding the fit statistics for the person results, none of the students presented an infit value greater than 1.5; only two third graders (0.8 % of the third grade sample) obtained outfit values greater than 1.5. Point-measure correlations in the TRC-n-3 ranged from -.28 for item AN12^{CC} to .55 for item TT8^{R}. Negative point-measure correlation coefficients were found for items RB3^{IC} and AN12^{CC} on the TRC-n-3. For these items and items AN3.15^{IC} and LB3.8^{R}, the highest mean ability value was not obtained by the group of students who chose the correct answer option.

Items of inadequate difficulty for third graders, fit statistics greater than 1.5, negative point-measure correlations and/or problems in the answer options (four items) were removed from TRC-n-3 and the test form, with 43 items, was recalibrated and linked to the TRC-n-2 (34 items). Table 3 shows the psychometric properties of the TRC-n-3 items. All the items revealed adequate psychometric characteristics in this new analysis. Fit statistics for the person results indicated that none of the students presented an infit value greater than 1.5 and only three third graders (1.3 % of the third grade sample) obtained outfit values greater than 1.5. Based on this recalibration of the TRC-n-3, displacement estimates were analyzed to evaluate the stability of the common items' difficulty between

Table 2 Psychometric properties of the items of the TRC-n-2 (34 items)

Item	b_i	SE	Infit	Outfit	R
RB1^{LC}	−0.10	0.14	1.00	0.98	.35
RB2^{LC}	−0.67	0.15	0.93	0.93	.39
RB5^{IC}	0.58	0.14	1.06	1.06	.29
RB3.6^{IC}	−0.32	0.15	1.07	1.18	.24
RB3.10^{IC}	0.62	0.15	1.02	1.04	.32
RB2.9 ^{R}	0.63	0.14	1.07	1.08	.26
RB2.10^{LC}	−0.98	0.16	0.83	0.73	.51
RB2.11^{IC}	0.78	0.15	0.90	0.87	.47
RB2.13^{R}	−0.43	0.15	0.94	0.92	.40
AN1^{IC}	0.40	0.14	1.06	1.06	.28
AN7^{IC}	0.14	0.14	1.07	1.08	.26
AN10^{CC}	−1.00	0.16	0.95	0.85	.37
AN3.9^{CC}	−0.19	0.14	0.94	0.93	.41
AN3.11^{IC}	0.70	0.15	1.05	1.08	.29
AN2.7^{IC}	0.33	0.14	1.00	1.01	.35
AN2.8^{IC}	0.89	0.15	1.14	1.16	.18
AN2.9^{IC}	1.12	0.15	1.11	1.17	.19
AN2.10^{R}	−1.07	0.16	0.94	0.83	.38
AN2.11^{IC}	0.04	0.14	0.97	0.96	.38
AN2.12^{LC}	0.52	0.14	0.88	0.87	.49
LB1^{LC}	0.18	0.14	1.04	1.04	.31
LB4^{IC}	1.45	0.16	1.19	1.41	.07
LB3.6^{IC}	0.15	0.14	1.07	1.08	.26
LB3.7^{IC}	−0.66	0.15	1.04	1.04	.27
LB2.6^{IC}	−0.96	0.16	0.98	1.05	.31
TT2.1^{LC}	−1.13	0.17	0.98	1.09	.28
TT1^{LC}	−0.49	0.15	0.96	0.93	.38
TT3^{LC}	−0.65	0.15	0.87	0.79	.48
TT6^{IC}	0.73	0.15	1.20	1.28	.10
TT2.5^{IC}	0.30	0.15	0.93	0.91	.44
TT2.6^{LC}	−1.19	0.17	0.91	0.80	.40
TT2.7^{IC}	0.06	0.15	1.02	1.04	.32
TT2.8^{IC}	−0.12	0.15	0.89	0.85	.48
TT2.9^{IC}	0.36	0.15	0.91	0.90	.46

Note. b_i item difficulty, *SE* standard error. Items are identified by the text to which they are related (*RB* Rita and Bruno, *AN* the animals, *LB* lost bread, *TT* two twins almost alike), followed by a number. The comprehension level assessed by each item is presented in superscript (*LC* literal comprehension, *IC* inferential comprehension, *R* reorganization, *CC* critical comprehension). Common items to the TRC-n-3 appear in bold

Table 3 Psychometric properties of the items of the TRC-n-3 (43 items)

Item	b_i	SE	Infit	Outfit	r	Displace
RB1^{LC}	−0.10^{A}	0.15	0.92	0.83	.49	0.14
RB2^{LC}	−0.67^{A}	0.17	1.14	1.33	.21	0.14
RB4^{IC}	1.07	0.14	1.15	1.19	.20	0.00
RB5^{IC}	0.58^{A}	0.14	1.20	1.35	.15	0.19
RB3.6^{IC}	−0.32^{A}	0.16	1.27	1.23	.37	0.58
RB3.7^{IC}	−1.26	0.21	0.95	0.79	.32	0.00
RB3.8^{IC}	−1.10	0.19	0.91	0.67	.40	0.00
RB3.9^{LC}	−1.35	0.21	0.97	0.77	.29	0.00
RB3.10^{IC}	0.62^{A}	0.14	1.03	1.05	.40	0.42
AN1^{IC}	0.40^{A}	0.14	0.80	0.73	.50	−0.44
AN6^{IC}	1.86	0.15	1.16	1.26	.18	0.00
AN7^{IC}	0.14^{A}	0.15	0.86	0.78	.46	−0.15
AN9^{LC}	0.48	0.14	0.96	0.93	.41	0.00
AN10^{CC}	−1.00^{A}	0.19	0.58	0.45	.19	−1.07
AN11^{IC}	1.09	0.14	1.11	1.12	.25	0.00
AN16^{R}	2.76	0.18	1.02	1.12	.28	0.00
AN3.9^{CC}	−0.19^{A}	0.16	1.26	1.29	.34	0.56
AN3.10^{IC}	0.24	0.15	1.08	1.13	.25	0.00
AN3.11^{IC}	0.70^{A}	0.14	0.88	0.85	.49	−0.07
AN3.12^{IC}	0.24	0.15	1.02	0.97	.34	0.00
AN3.13^{IC}	−0.54	0.17	1.05	1.07	.24	0.00
AN3.14^{IC}	1.11	0.14	1.19	1.24	.16	0.00
AN3.16^{CC}	0.40	0.14	1.18	1.25	.15	0.00
AN3.17^{IC}	0.16	0.15	1.00	1.07	.33	0.00
LB1^{LC}	0.18^{A}	0.15	0.96	0.89	.39	−0.04
LB3^{CC}	−0.69	0.18	0.95	0.88	.35	0.00
LB4^{IC}	1.45^{A}	0.14	1.13	1.16	.32	−0.37
LB6^{IC}	0.43	0.14	1.03	1.02	.34	0.00
LB3.5^{CC}	1.50	0.14	1.06	1.05	.32	0.00
LB3.6^{IC}	0.15^{A}	0.15	0.79	0.73	.39	−0.73
LB3.7^{IC}	−0.66^{A}	0.17	1.26	1.40	.18	0.27
LB13^{IC}	0.88	0.14	1.03	1.04	.34	0.00
LB15^{CC}	0.99	0.14	1.13	1.18	.22	0.00
TT1^{LC}	−0.49^{A}	0.17	0.92	0.93	.40	0.04
TT3^{LC}	−0.65^{A}	0.17	0.69	0.51	.52	−0.25
TT6^{IC}	0.73^{A}	0.14	1.00	1.03	.38	0.14
TT7^{LC}	−0.14	0.15	0.94	0.85	.41	0.00
TT8^{R}	0.55	0.14	0.83	0.77	.56	0.00
TT9^{LC}	−0.03	0.15	0.85	0.77	.51	0.00
TT10^{LC}	0.04	0.15	0.93	0.88	.42	0.00
TT11^{LC}	0.21	0.15	0.94	0.93	.42	0.00
TT12^{IC}	0.27	0.15	0.93	0.91	.43	0.00
TT3.10^{IC}	−0.97	0.19	0.99	0.91	.28	0.00

Note. b_i item difficulty, *SE* standard error. Items are identified by the text to which they are related (*RB* Rita and Bruno, *AN* the animals, *LB* lost bread, *TT* two twins almost alike), followed by a number. The comprehension level assessed by each item is presented in superscript (*LC* literal comprehension, *IC* inferential comprehension, *R* reorganization, *CC* critical comprehension). Common items to the TRC-n-2 and TRC-n-4 appear in bold. ^{A} = Anchor item with a fixed value

the second and the third grade test forms. From the 17 common items between the TRC-n-2 and TRC-n-3, two items presented displacement values far from the reference value of 0.5: AN10^{CC} (−1.07) and LB3.6^{IC} (−0.73). These items did not meet the criteria of maintaining reasonably stable difficulty across test forms so they were not suitable to work as anchor items between the TRC-n-2 and TRC-n-3. Displacement results of the remaining common items varied between −0.44, in the item AN1^{IC}, and 0.58, in the item RB3.6^{IC} (Table 3). After considering this analysis and evaluating the distribution of the remaining common items along the difficulty range of the test and the items' representativeness of the test content, six items were selected to equate the TRC-n-2 and TRC-n-3 test forms: RB1^{LC}, RB2^{LC}, RB3.6^{IC}, RB3.10^{IC}, AN3.9^{CC} and TT6^{IC}.

After this procedure, new computer runs were performed to calibrate and link the TRC-n-4 initial pool of items to the TRC-n-3 with 43 items. Item analysis of the TRC-n-4 showed that item difficulty ranged from −1.00 to 3.25 and person ability ranged from −1.31 to 3.69. None of the items was identified as too easy or too difficult for the fourth grade. Item AN16^{R} exceeded the reference value of 1.5 for infit. This item, along with items LB9^{R} and LB10^{IC}, had outfit values higher than 1.5. Regarding the fit statistics for the person results in the TRC-n-4, none of the students presented an infit value greater than 1.5 and only nine fourth graders (3.9 % of the fourth grade sample) obtained outfit values greater than 1.5. Point-measure correlations in the TRC-n-4 ranged between -.12 for item AN12^{CC} and .56 for item TT11^{LC}. Negative point-measure correlation coefficients were found for items RB3^{IC}, AN12^{CC} and LB10^{IC}. For these items and items LB2^{IC} and LB9^{R}, the highest mean ability value was not obtained by the group of students who chose the correct answer option.

Items of inadequate difficulty, fit statistics greater than 1.5, negative point-measure correlations and/or problems in the answer options (six items) were removed from TRC-n-4 and this test form with 51 items was recalibrated and again linked to the TRC-n-3. Table 4 shows the psychometric properties of the items of the TRC-n-4. None of the items revealed inappropriate

Table 4 Psychometric properties of the items of the TRC-n-4 (51 items)

Item	b_i	SE	Infit	Outfit	r	Displace
RB1^{LC}	−0.10^{A}	0.16	0.68	0.58	.39	−0.64
RB2^{LC}	−0.67^{A}	0.19	1.25	1.18	.30	0.36
RB4^{IC}	1.07^{A}	0.14	1.08	1.09	.28	0.02
RB5^{IC}	0.58^{A}	0.15	1.11	1.14	.28	0.18
RB6^{R}	0.27	0.15	0.93	0.95	.40	0.00
RB7^{IC}	0.07	0.16	0.92	0.81	.43	0.00
RB8^{IC}	0.32	0.15	1.01	0.97	.34	0.00
RB9^{IC}	1.87	0.15	1.28	1.44	.04	0.00
RB10^{IC}	0.72	0.15	1.13	1.18	.21	0.00
RB11^{IC}	1.98	0.15	1.04	1.10	.31	0.00
AN1^{IC}	0.40^{A}	0.15	0.84	0.80	.37	−0.47
AN2^{IC}	0.97	0.14	0.97	0.94	.40	0.00
AN3^{LC}	0.71	0.15	1.00	0.96	.37	0.00
AN4^{IC}	1.04	0.14	0.93	0.90	.44	0.00
AN5^{IC}	0.99	0.14	0.91	0.87	.47	0.00
AN6^{IC}	1.86^{A}	0.15	1.32	1.39	.24	−0.78
AN7^{IC}	0.14^{A}	0.16	0.80	0.75	.41	−0.36
AN8^{CC}	0.33	0.15	0.95	0.93	.40	0.00
AN9^{LC}	0.48^{A}	0.15	1.22	1.27	.37	0.69
AN10^{CC}	−1.00^{A}	0.21	0.60	0.51	.26	−0.71
AN11^{IC}	1.09^{A}	0.14	1.14	1.14	.24	0.09
AN13^{IC}	0.94	0.14	1.07	1.11	.28	0.00
AN14^{IC}	1.31	0.14	1.07	1.07	.30	0.00
AN15^{LC}	0.74	0.15	1.14	1.26	.19	0.00
AN17^{CC}	0.19	0.16	1.06	1.18	.25	0.00
LB1^{LC}	0.18^{A}	0.16	0.98	1.01	.32	−0.06
LB3^{CC}	−0.69^{A}	0.19	1.10	1.06	.35	0.24
LB4^{IC}	1.45^{A}	0.14	0.99	1.00	.41	−0.42
LB5^{IC}	−0.01	0.16	0.96	0.90	.38	0.00
LB6^{IC}	0.43^{A}	0.15	0.93	0.92	.35	−0.22
LB7^{CC}	1.56	0.14	1.09	1.12	.27	0.00
LB8^{IC}	1.19	0.14	0.99	0.98	.39	0.00
LB11^{LC}	1.21	0.14	0.99	0.96	.39	0.00
LB12^{IC}	1.26	0.14	1.11	1.11	.25	0.00
LB13^{IC}	0.88^{A}	0.14	1.04	1.01	.34	0.06
LB14^{IC}	2.78	0.17	1.07	1.22	.23	0.00
LB15^{CC}	0.99^{A}	0.14	1.23	1.27	.14	0.17
TT1^{LC}	−0.49^{A}	0.18	0.99	0.83	.45	0.20
TT2^{LC}	1.24	0.14	1.18	1.25	.16	0.00
TT3^{LC}	−0.65^{A}	0.19	0.71	0.51	.44	−0.29
TT4^{LC}	−0.90	0.21	0.93	0.70	.37	0.00
TT5^{IC}	−0.06	0.17	0.86	0.74	.48	0.00
TT6^{IC}	0.73^{A}	0.15	1.20	1.23	.31	0.63
TT7^{LC}	−0.14^{A}	0.17	1.08	0.99	.46	0.38
TT8^{R}	0.55^{A}	0.15	0.88	0.81	.55	0.19
TT9^{LC}	−0.03^{A}	0.17	0.82	0.73	.52	0.00
TT10^{LC}	0.04^{A}	0.16	0.93	0.87	.42	0.01
TT11^{LC}	0.21^{A}	0.16	0.78	0.66	.55	−0.08
TT12^{IC}	0.27^{A}	0.16	1.00	0.93	.39	0.08
TT13^{LC}	0.83	0.15	0.88	0.88	.49	0.00
TT14^{IC}	1.02	0.14	1.01	1.01	.36	0.00

Note. b_i item difficulty, *SE* standard error. Items are identified by the text to which they are related (*RB* Rita and Bruno, *AN* the animals, *LB* lost bread, *TT* Two twins almost alike), followed by a number. The comprehension level assessed by each item is presented in superscript (*LC* literal comprehension, *IC* inferential comprehension, *R* reorganization, *CC* critical comprehension). Common items to the TRC-n-3 appear in bold. ^{A} = Anchor item with a fixed value

psychometric characteristics in this new analysis. Fit statistics for the person results improved since none of the students presented an infit value greater than 1.5 and only three fourth graders (1.3 % of the sample) obtaining outfit values greater than 1.5.

The following step was the analysis of the displacement values of the common items between the third and fourth grade test forms. From the remaining 25 common items, five items presented displacement values far from 0.5: RB1^{LC} (−0.64), AN6^{IC} (−0.78), AN9^{LC} (0.69), AN10^{CC} (−0.71) and G6^{IC} (0.63). Therefore, these items were not acceptable for use to equate the TRC-n-3 and TRC-n-4. Displacements values of the other items ranged from to −0.47, in AN1^{IC}, to 0.38 in G7^{LC} (see Table 4). Based on this analysis and on the remaining items' difficulty and comprehension level, six items were selected to equate the TRC-n-3 and TRC-n-4 test forms: RB2^{LC}, AN11^{IC}, LB13^{IC}, LB15^{CC}, TT8^{R} and TT10^{LC}.

After selecting the anchor items for each test form, unique items were selected to streamline the tests according to the criteria previously mentioned: the spread of difficulty, redundancy and the comprehension level that was assessed. When two or more items were of similar difficulty levels and they measured the same level of comprehension, the item with a higher point-measure correlation was retained. The number of items retained in each final version of the TRC-n forms is presented in Table 1 ("Selected items for the final test forms" column). The final test forms of the TRC-n were composed of 27 items each, six of which were anchor items between adjacent test forms (22.2 % of the total number of items).

A new set of calibrations was performed to link the final forms of each test. The item parameter locations on the TRC-n vertical scale are presented in Figs. 2, 3 and 4. The difficulty of the items ranged between −1.19

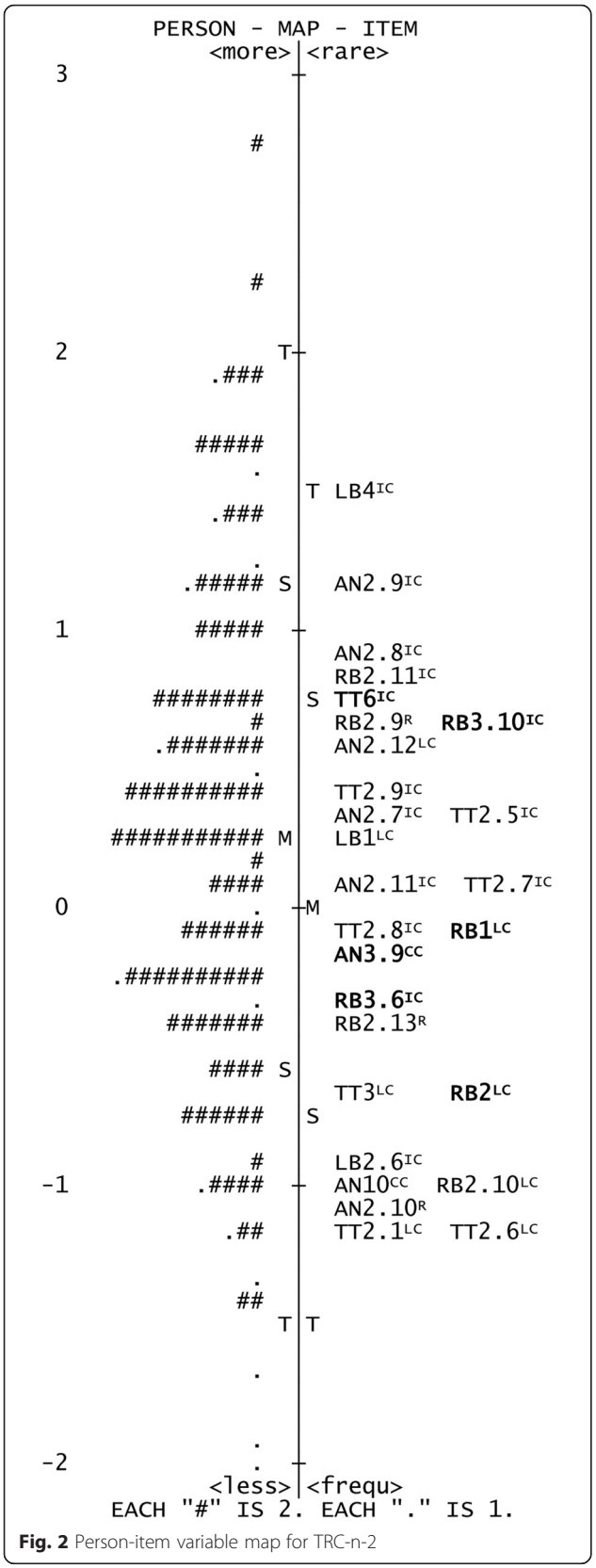

Fig. 2 Person-item variable map for TRC-n-2

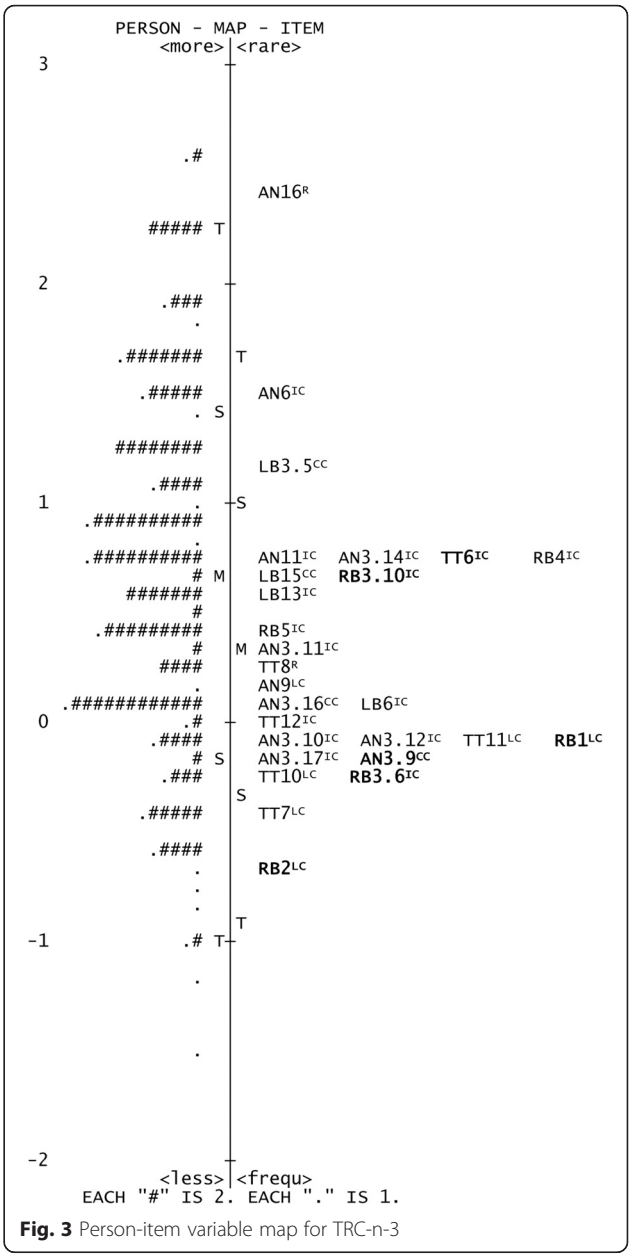

Fig. 3 Person-item variable map for TRC-n-3

and 1.51 for the TRC-n-2 (mean = 0.00; SD = 0.75), between −0.66 and 2.41 for the TRC-n-3 (mean = 0.34; SD = 0.65) and between −0.66 and 2.40 (mean = 0.44; SD = 0.64) for the TRC-n-4.

Items are identified by the text to which they are related (RB = Rita and Bruno; AN = The animals; LB = The lost bread; TT = Two twins almost alike), followed by the item's number. The comprehension level assessed by each item is presented in superscript (LC = literal comprehension; IC = inferential comprehension; R = reorganization; CC = critical comprehension). Common items appear in bold (Figs. 2 and 3 – common items between grades 2 and 3; Fig. 4 – common items between grades 3 and 4).

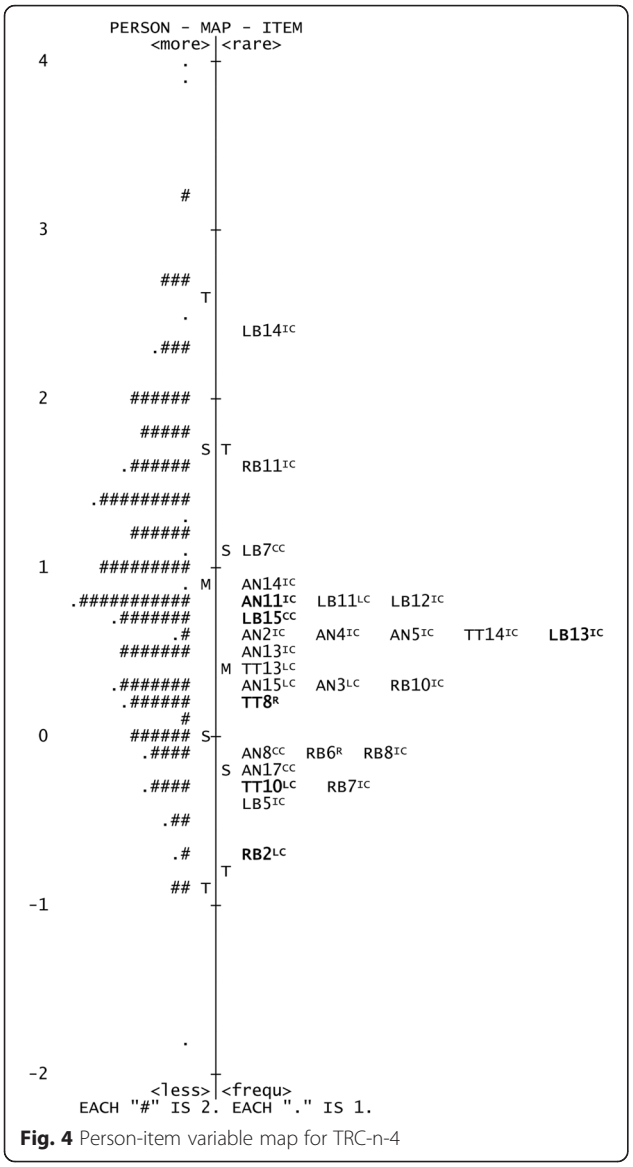

Fig. 4 Person-item variable map for TRC-n-4

None of the items yielded infit or outfit values that were greater than 1.5. After producing these final versions of the test forms, the person fit statistics were altered: 11 s graders (4.8 % of the second grade sample), three third graders (1.3 % of the third grade sample) and two fourth graders (0.9 % of the fourth grade sample) obtained outfit values higher than 1.5. Compared to the percentage of misfit students in the calibration of the initial pool of items, this calibration yielded increased percentages of second and third grade misfit students and a decreased percentage of misfit fourth grade students.

This final set of items yielded item difficulty levels that were more appropriate to each sample group. The lowest item difficulty value was higher than the lowest person ability value for all three test forms (−1.92, −1.54, and −1.78 for second, third and fourth graders, respectively). Additionally, the highest item difficulty value was lower than the highest person ability value (2.76, 2.59 and 3.91 for second, third, and fourth graders, respectively). Mean values of the person ability scaled scores for the TRC-n were 0.26 (SD = .87) for the TRC-n-2, 0.65 (SD = .81) for the TRC-n-3 and 0.87 (SD = .87) for the TRC-n-4. As grades increased, person ability values were significantly greater, $F(2, 690) = 33.019$, $p < .001$. Bonferroni post-hoc tests yielded statistically significant differences ($p < .05$) between the mean results of each grade.

Results of the PCA of the residuals of the TRC-n final test forms revealed that all the secondary dimensions had eigenvalues less than 2.0. The correlations of the residuals of the items ranged from zero to .27. Regarding the reliability of the TRC-n, the PSR and KR20 coefficients were moderate (PSR: TRC-n-2 = .70, TRC-n-3 = .69, TRC-n-4 = .72; KR20: TRC-n-2 = .77, TRC-n-3 = .72, TRC-n-4 = .76) and the ISR coefficients were very high (TRC-n-2 = .96, TRC-n-3 = .95, TRC-n-4 = .94).

Test of reading comprehension of expository texts (TRC-e)

On the TRC-e-2 initial pool of items, item difficulty ranged from −1.95 to 2.70 and person ability ranged from −1.80 to 2.67. The minimum value for person ability exceeded the minimum value for item difficulty, indicating the presence of items excessively easy for this grade. This was the case of item $FN1^{LC}$. The maximum item difficulty value was greater than the maximum person ability value since item $KS20^{R}$ was excessively difficult for the second graders. None of the items had an infit value greater than 1.5 and only item $KS20^{R}$ had an outfit value greater than 1.5. Regarding the fit statistics for the person results on the TRC-e forms, none of the students presented an infit value greater than 1.5 and only nine second graders (3.6 % of the sample), exhibited outfit values that exceeded the reference value. Point-measure correlations ranged from -.05 for item $KS20^{R}$ to .48 for items $C3.6^{LC}$ and $BE2.2^{IC}$. Item $KS20^{R}$ was the only item exhibiting a negative point-measure correlation coefficient. This item, along with item $BE2.6^{IC}$, also presented problems with the quality of answer options, suggesting that the greatest mean ability value was not observed for the participants who chose the correct option for each of these items.

In the TRC-e-2, three items were removed as a result of the item analysis. The test form with 33 selected items was recalibrated. Table 5 shows the psychometric properties of the items of the TRC-e-2. None of the items revealed inappropriate psychometric characteristics in this new analysis and, regarding the fit statistics for the person results, none of the students presented an outfit value that exceeded 1.5 and only one student (0.4 % of the sample) presented an outfit value above the reference value.

Table 5 Psychometric properties of the items of the TRC-e-2 (33 items)

Item	b_i	SE	Infit	Outfit	r
FN2^{LC}	−1.71	0.18	0.96	0.92	.30
FN3.3^{CC}	0.22	0.14	1.00	0.98	.34
FN2.4^{LC}	−0.76	0.14	1.06	1.21	.21
FN2.5^{IC}	−0.24	0.14	1.09	1.20	.20
FN2.6^{IC}	1.05	0.15	1.18	1.38	.07
C2.1^{LC}	−0.48	0.14	1.08	1.16	.20
C2.2^{LC}	−0.83	0.15	0.89	0.83	.45
C2^{IC}	0.38	0.14	1.07	1.09	.24
C2.4^{IC}	−0.24	0.14	1.08	1.12	.22
C4^{LC}	−0.49	0.14	0.92	0.89	.42
C3.5^{IC}	−0.41	0.14	0.90	0.86	.45
C3.6^{LC}	0.26	0.14	0.89	0.86	.48
C19^{IC}	−0.67	0.14	0.91	0.88	.42
BE2.1^{R}	−0.42	0.14	0.92	0.88	.43
BE2.2^{IC}	−1.07	0.15	0.86	0.76	.47
BE2.3^{LC}	0.24	0.13	0.92	0.90	.44
BE2.4^{CC}	0.85	0.14	0.99	1.01	.33
BE2.5^{LC}	0.60	0.14	1.10	1.13	.20
BE2.7^{R}	1.02	0.14	1.16	1.29	.09
BE7^{IC}	0.31	0.14	0.97	0.96	.38
BE9^{IC}	0.77	0.14	1.12	1.14	.18
BE12^{LC}	−0.15	0.14	0.91	0.88	.44
BE13^{LC}	−0.04	0.14	0.96	0.96	.38
BE14^{IC}	0.49	0.14	1.05	1.06	.27
BE15^{LC}	−0.08	0.14	0.98	0.95	.36
KS11^{IC}	0.14	0.13	1.08	1.07	.24
KS12^{LC}	−0.52	0.14	0.92	0.87	.42
KS13^{LC}	1.22	0.15	0.95	0.95	.37
KS14^{IC}	0.31	0.14	0.95	0.93	.40
KS15^{IC}	−0.14	0.14	0.99	0.99	.34
KS16^{LC}	0.54	0.14	1.19	1.19	.10
KS18^{LC}	−0.05	0.14	0.89	0.86	.47
KS2.9^{IC}	−0.08	0.14	1.00	1.01	.33

Note. b_i item difficulty, *SE* standard error. Items are identified by the text to which they are related (*FN* fireflies' night, *C* the caravels, *BE* observing birds in the estuary, *KS* the king Sebastian), followed by a number. The comprehension level assessed by each item is presented in superscript (*LC* literal comprehension, *IC* inferential comprehension, *R* reorganization, *CC* critical comprehension). Common items to the TRC-e-3 appear in bold

The TRC-e-3 initial pool of items was then linked to the TRC-e-2 (33 items). On the TRC-e-3, item difficulty ranged from −1.88 to 2.41 and person ability ranged from −0.98 to 3.56. The minimum values for person ability exceeded the minimum values for item difficulty, indicating that some items were excessively easy, namely items FN1^{LC}, FN2^{LC}, FN7^{IC}, BE1^{LC} and KS3^{LC}. The maximum item difficulty value was lower than the maximum person ability value in the TRC-e-3, meaning that none of the items was excessively difficult for the third graders. None of the items had infit or outfit values greater than 1.5. Regarding the fit statistics for the person results, none of the students presented an infit value greater than 1.5 and only eight third graders (3.6 % of the sample) exhibited outfit values that exceeded the reference value. Point-measure correlations ranged from -.15 for item C20^{LC} to .52 for item C2^{IC}. Item C20^{LC} was the only item to present a negative point-measure correlation coefficient and the only one where the greatest mean ability value was not observed for the participants who chose the correct alternative.

From this test form, six items were removed as they presented inadequate difficulty for the third grade, negative point-measure correlations and/or problems in the answer options. The test form with 47 items was recalibrated and linked to the TRC-e-2. Table 6 shows the psychometric properties of the items of the TRC-e-3. All the items revealed adequate psychometric characteristics in this new analysis. Fit statistics for the person results indicated that none of the students presented an infit value greater than 1.5 and only one (0.5 % of the sample) obtained outfit values greater than 1.5. Based on this recalibration of the TRC-e-3, displacement estimates were analyzed to evaluate the stability of the common items' difficulty between the second and the third grade test forms. From the 17 common items between the TRC-e-2 and TRC-e-3, seven items presented displacement values far from the reference value of 0.5: FN3.3^{CC} (−0.66), C4^{LC} (−0.80), BE14^{IC} (−0.86), BE15^{LC} (−0.60), KS11^{IC} (1.00), KS12^{LC} (−0.75) and KS13^{LC} (−0.74). Therefore, these common items were not adequate to equate the second and third grade test forms. Displacement results of the remaining items varied between −0.58, in the item BE13^{LC}, and 0.40, in the item KS16^{IC} (see Table 6). According to the items' displacement values, difficulty and comprehension level, eight items were retained as anchor items on the TRC-e-2 and the TRC-e-3: C2^{IC}, C3.5^{IC}, C3.6^{LC}, C19^{IC}, BE9^{IC}, BE12^{LC}, KS14^{IC}, and KS15^{IC}.

New computer runs were then performed to calibrate and link the TRC-e-4 initial pool of items to the TRC-e-3 with 47 items. On the TRC-e-4, item difficulty ranged from −1.69 to 2.46 and person ability ranged from −1.10 to 3.04. The minimum values for person ability exceeded the minimum values for item difficulty, indicating that some items were excessively easy, such as the following items: FN1^{LC}, FN2^{LC}, FN7^{IC}, BE1^{LC}, KS1^{R}, KS3^{LC}. The maximum item difficulty value was lower than the maximum person ability value in the TRC-e-4, meaning that none of the items was excessively difficult for the fourth

Table 6 Psychometric properties of the items of the TRC-e-3 (47 items)

Item	b_i	SE	Infit	Outfit	R	Displace
FN3.3^{CC}	0.22^{A}	0.15	0.87	0.80	.38	−0.66
FN3^{IC}	1.03	0.15	1.20	1.25	.14	0.00
FN4^{IC}	−0.69	0.17	0.89	0.72	.43	0.00
FN5^{LC}	1.08	0.15	1.16	1.22	.18	0.00
FN6^{IC}	0.99	0.15	1.08	1.09	.28	0.00
C2^{IC}	0.38^{A}	0.15	0.83	0.78	.52	−0.25
C3.2^{LC}	−0.26	0.15	1.02	1.03	.28	0.00
C3.3^{IC}	−0.65	0.17	1.08	1.41	.12	0.00
C4^{LC}	−0.49^{A}	0.16	0.64	0.54	.39	−0.80
C3.5^{IC}	−0.41^{A}	0.16	0.88	0.75	.43	−0.06
C3.6^{LC}	0.26^{A}	0.15	0.98	0.93	.42	0.17
C3.7^{LC}	−0.97	0.18	0.97	0.83	.31	0.00
C3.8^{IC}	−0.40	0.16	0.91	0.84	.41	0.00
C10^{LC}	1.82	0.16	1.11	1.12	.24	0.00
C19^{IC}	−0.67^{A}	0.17	0.90	0.84	.38	−0.05
BE2^{LC}	0.01	0.15	0.98	0.98	.35	0.00
BE3^{IC}	0.16	0.15	0.97	0.92	.38	0.00
BE4^{IC}	0.06	0.15	0.99	0.99	.34	0.00
BE5^{IC}	−0.49	0.16	1.12	1.23	.14	0.00
BE6^{IC}	0.74	0.15	1.07	1.07	.29	0.00
BE8^{R}	0.12	0.15	0.97	0.92	.38	0.00
BE3.8^{R}	−0.63	0.17	0.90	0.75	.42	0.00
BE9^{IC}	0.77^{A}	0.15	1.04	1.03	.32	0.17
BE11^{LC}	−0.36	0.16	1.04	0.99	.27	0.00
BE12^{LC}	−0.15^{A}	0.15	0.97	0.88	.49	0.30
BE13^{LC}	−0.04^{A}	0.15	0.77	0.68	.44	−0.58
BE14^{IC}	0.49^{A}	0.15	0.91	0.87	.40	−0.86
BE15^{LC}	−0.08^{A}	0.15	0.70	0.61	.51	−0.60
BE16^{LC}	1.01	0.15	0.99	0.99	.38	0.00
BE17^{IC}	0.51	0.15	1.03	0.99	.33	0.00
BE3.17^{R}	−0.04	0.15	1.08	1.17	.21	0.00
KS3.1^{LC}	−0.24	0.16	1.05	1.07	.24	0.00
KS4^{IC}	−0.38	0.16	0.98	0.95	.32	0.00
KS5^{IC}	0.28	0.15	1.06	1.06	.28	0.00
KS6^{IC}	−0.51	0.16	1.00	1.01	.29	0.00
KS8^{IC}	−0.69	0.17	0.96	0.88	.34	0.00
KS9^{IC}	0.18	0.15	0.87	0.82	.49	0.00
KS10^{CC}	0.71	0.15	0.92	0.90	.45	0.00
KS11^{IC}	0.14^{A}	0.15	1.21	1.27	.40	1.00
KS12^{LC}	−0.52^{A}	0.16	1.38	1.35	.35	0.75
KS13^{LC}	1.22^{A}	0.15	1.01	1.04	.21	0.74
KS14^{IC}	0.31^{A}	0.15	0.99	1.00	.33	−0.09
KS15^{IC}	−0.14^{A}	0.15	0.98	0.91	.40	0.11
KS16^{LC}	0.54^{A}	0.15	1.24	1.29	.12	0.40
KS19^{IC}	0.27	0.15	1.01	1.01	.33	0.00
KS20^{R}	2.46	0.19	1.15	1.40	.11	0.00
KS23^{R}	0.76	0.15	0.96	0.95	.41	0.00

Note. b_i item difficulty, *SE* standard error. Items are identified by the text to which they are related (FN fireflies' night, *C* the caravels, *BE* observing birds in the estuary, *KS* the king Sebastian), followed by a number. The comprehension level assessed by each item is presented in superscript (*LC* literal comprehension, *IC* inferential comprehension, *R* reorganization, *CC* critical comprehension). Common items to the TRC-e-2 and TRC-e-4 appear in bold. ^{A} = Anchor item with a fixed value

graders. Only item KS11^{IC} presented an infit value greater than 1.5. This item and item BE18^{R} exhibited outfit values greater than 1.5. Regarding the fit statistics for the person results on the TRC-e forms, none of the students presented an infit value greater than 1.5 and only seven fourth graders (2.6 % of the sample) exhibited outfit values that exceeded the reference value. Point-measure correlations in the TRC-e-4 ranged between -.21 for item KS20^{R} and .51 for item BE13^{LC}. Negative point-measure correlation coefficients were obtained for items BE18^{R} and KS20^{R}. These items plus item FN11^{R} presented problems with the quality of answer options.

This item analysis resulted in the exclusion of 10 items from the TRC-e-4 due to their inadequate psychometric characteristics. The TRC-e-4 test form with 69 items was recalibrated and again linked to the TRC-e-3. Table 7 shows the psychometric properties of the items of the TRC-e-4. None of the items revealed inappropriate psychometric characteristics in this new analysis. Item FN12^{CC} was the only item to exhibit an outfit value higher than 1.5 (1.53). Fit statistics for the person results improved with none of the students presenting an infit value greater than 1.5 and only one fourth grader (0.4 % of the sample) obtaining outfit values greater than 1.5. The analysis of the displacement values of the 34 common items between the third and fourth grade test forms evidenced that seven items presented displacement values higher than 0.5: C2^{IC} (−0.64), BE12^{LC} (0.84), BE14^{IC} (−0.87), KS6^{IC} (−0.83), KS8^{IC} (−0.64), KS14^{IC} (−0.89) and KS16^{LC} (0.93). These items, therefore, were not suitable to be used as anchor items. Displacements values of the remaining items ranged from to −0.38, in C10^{LC}, to 0.40 in BE17^{IC} (Table 7). From these items, eight were retained as anchor items in the final versions of the TRC-e-3 and the TRC-e-4, according to their difficulty and the comprehension level assessed by these questions: FN5^{LC}, BE8^{R}, BE11^{LC}, BE16^{LC}, BE17^{IC}, KS5^{IC}, KS10^{CC} and KS23^{R}.

After selecting the anchor items for each TRC-e form, unique items' selection was performed to streamline the tests according to the difficulty of the items, their

Table 7 Psychometric properties of the items of the TRC-e-4 (69 items)

Item	b_i	SE	Infit	Outfit	r	Displace
FN3^{IC}	1.03^{A}	0.14	1.13	1.21	.25	0.09
FN4^{IC}	−0.69^{A}	0.18	0.83	0.70	.37	−0.20
FN5^{LC}	1.08^{A}	0.14	1.22	1.27	.17	0.14
FN6^{IC}	0.99^{A}	0.14	1.13	1.12	.27	0.02
FN8^{LC}	−0.74	0.18	0.88	0.69	.44	0.00
FN9^{LC}	0.34	0.14	0.96	0.94	.41	0.00
FN10^{R}	0.84	0.14	1.01	0.99	.38	0.00
FN12^{CC}	2.05	0.14	1.31	1.53	.03	0.00
C1^{IC}	0.16	0.15	1.01	0.98	.35	0.00
C2^{IC}	0.38^{A}	0.14	0.80	0.75	.42	−0.65
C3^{R}	0.37	0.14	0.98	0.94	.40	0.00
C4^{LC}	−0.49^{A}	0.17	0.94	0.81	.43	0.08
C5^{IC}	−0.58	0.17	0.85	0.62	.49	0.00
C6^{IC}	−0.83	0.18	0.97	1.03	.29	0.00
C7^{LC}	1.05	0.13	1.00	1.02	.39	0.00
C8^{IC}	0.68	0.14	1.09	1.09	.29	0.00
C9^{R}	0.99	0.14	0.99	0.99	.40	0.00
C10^{LC}	1.82^{A}	0.14	1.32	1.43	.15	−0.39
C11^{IC}	1.76	0.14	1.15	1.23	.22	0.00
C12^{LC}	−0.57	0.17	0.99	0.96	.31	0.00
C13^{CC}	1.17	0.14	1.02	1.04	.36	0.00
C14^{R}	0.03	0.15	1.00	1.01	.34	0.00
C15^{IC}	0.55	0.14	1.02	1.03	.35	0.00
C16^{R}	−0.31	0.16	0.86	0.79	.47	0.00
C17^{IC}	−0.51	0.17	0.89	0.81	.42	0.00
C18^{IC}	1.15	0.13	0.98	0.97	.42	0.00
C19^{IC}	−0.67^{A}	0.18	0.74	0.69	.42	−0.25
C20^{LC}	0.03	0.15	0.95	0.88	.41	0.00
BE2^{LC}	0.01^{A}	0.15	0.91	0.84	.46	0.03
BE3^{IC}	0.16^{A}	0.15	1.00	0.97	.41	0.14
BE4^{IC}	0.06^{A}	0.15	0.93	0.86	.35	−0.18
BE5^{IC}	−0.49^{A}	0.17	0.99	1.01	.31	0.01
BE6^{IC}	0.74^{A}	0.14	1.07	1.08	.35	0.22
BE7^{IC}	0.56	0.14	1.03	0.99	.35	0.00
BE8^{R}	0.12^{A}	0.15	1.06	1.05	.38	0.20
BE9^{IC}	0.77^{A}	0.14	1.04	1.06	.40	0.35
BE10^{IC}	0.07	0.15	0.89	0.79	.48	0.00
BE11^{LC}	−0.36^{A}	0.16	1.09	1.05	.31	0.14
BE12^{LC}	−0.15^{A}	0.16	1.32	1.34	.50	0.83
BE13^{LC}	−0.04^{A}	0.15	0.73	0.65	.50	−0.29
BE14^{IC}	0.49^{A}	0.14	0.75	0.67	.46	−0.88
BE15^{LC}	−0.08^{A}	0.15	0.85	0.85	.34	−0.34
BE16^{LC}	1.01^{A}	0.14	0.96	0.96	.43	0.07
BE17^{IC}	0.51^{A}	0.14	1.09	1.06	.41	0.39
BE19^{LC}	0.95	0.14	1.08	1.07	.30	0.00
BE20^{IC}	−0.48	0.17	1.00	1.09	.29	0.00
BE21^{IC}	1.27	0.14	1.16	1.19	.22	0.00
BE22^{LC}	0.48	0.14	1.11	1.10	.26	0.00
KS2^{IC}	0.87	0.14	1.04	1.03	.35	0.00
KS4^{IC}	−0.38^{A}	0.17	1.13	1.08	.34	0.23
KS5^{IC}	0.28^{A}	0.15	1.09	1.09	.34	0.19
KS6^{IC}	−0.51^{A}	0.18	0.66	0.62	.26	−0.84
KS7^{IC}	−0.20	0.16	0.97	0.84	.38	0.00
KS8^{IC}	−0.69^{A}	0.19	0.64	0.55	.33	−0.65
KS9^{IC}	0.18^{A}	0.15	0.88	0.78	.38	−0.33
KS10^{CC}	0.71^{A}	0.14	1.14	1.16	.33	0.42
KS12^{LC}	−0.52^{A}	0.18	0.99	0.88	.37	0.06
KS13^{LC}	1.22^{A}	0.14	1.15	1.25	.22	0.31
KS14^{IC}	0.31^{A}	0.15	0.72	0.62	.44	−0.90
KS15^{IC}	−0.14^{A}	0.16	0.90	0.84	.39	−0.12
KS16^{LC}	0.54^{A}	0.15	1.27	1.30	.39	0.91
KS17^{IC}	0.56	0.15	0.91	0.83	.49	0.00
KS18^{LC}	0.34	0.15	0.93	0.85	.45	0.00
KS19^{IC}	0.27^{A}	0.15	0.96	0.99	.38	−0.05
KS21^{IC}	−0.82	0.19	0.97	0.95	.29	0.00
KS22^{IC}	0.57	0.15	0.99	0.97	.38	0.00
KS23^{R}	0.76^{A}	0.14	1.01	1.06	.35	−0.09
KS24^{IC}	0.62	0.15	0.92	0.88	.47	0.00
KS25^{IC}	0.95	0.14	0.92	0.91	.47	0.00

Note. b_i item difficulty, *SE* standard error. Items are identified by the text to which they are related (FN fireflies' night, *C* the caravels, *BE* observing birds in the estuary, *KS* the king Sebastian), followed by a number. The comprehension level assessed by each item is presented in superscript (*LC* literal comprehension, *IC* inferential comprehension, *R* reorganization, *CC* critical comprehension). Common items to the TRC-e-3 appear in bold. ^{A} = Anchor item with a fixed value

redundancy and their comprehension level. The higher point-measure correlation criterion was considered when two or more items were of similar difficulty levels and they measured the same level of comprehension. The number of items retained in each final version of the TRC-e forms is presented in Table 1 (column "Selected items for the final test forms"). Each final version of the TRC-e was composed of 33 items, eight of which were anchor items (24.2 % of the total number of items).

The item parameter locations on the TRC-e vertical scale are presented in Figs. 5, 6 and 7. Regarding TRC-e, items' difficulty values ranged between −1.71 and 1.22 for the TRC-e-2 (mean = 0.00; SD = 0.64), between −1.04 and 2.47 for the TRC-e-3 (mean = 0.19; SD = 0.73) and between −0.81 and 1.86 (mean = 0.33; SD =

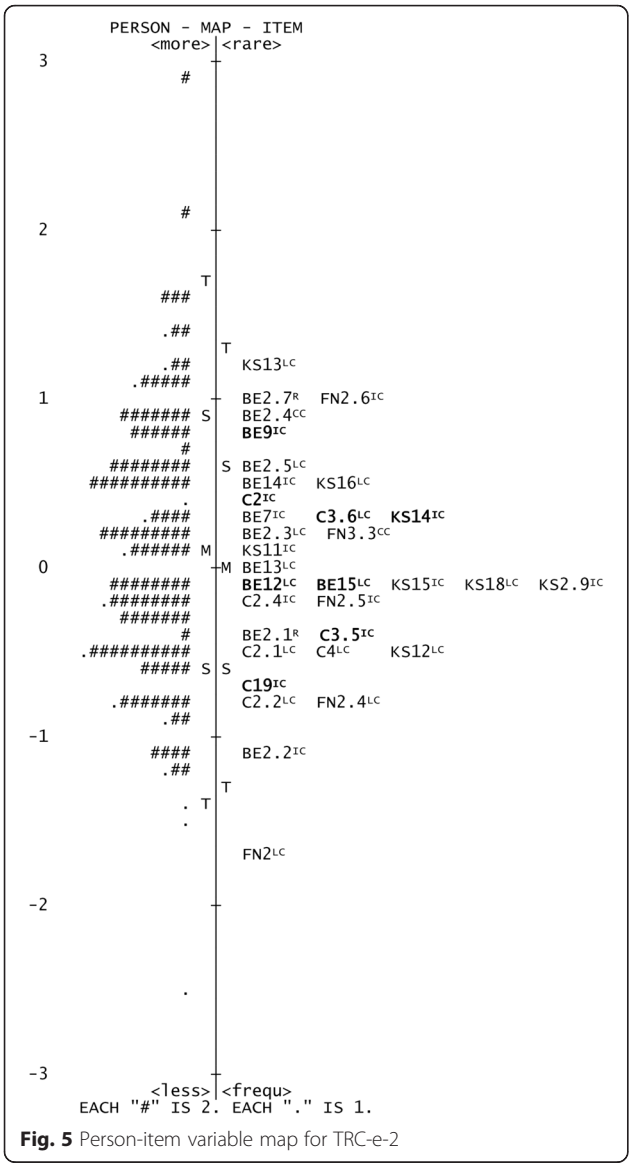

Fig. 5 Person-item variable map for TRC-e-2

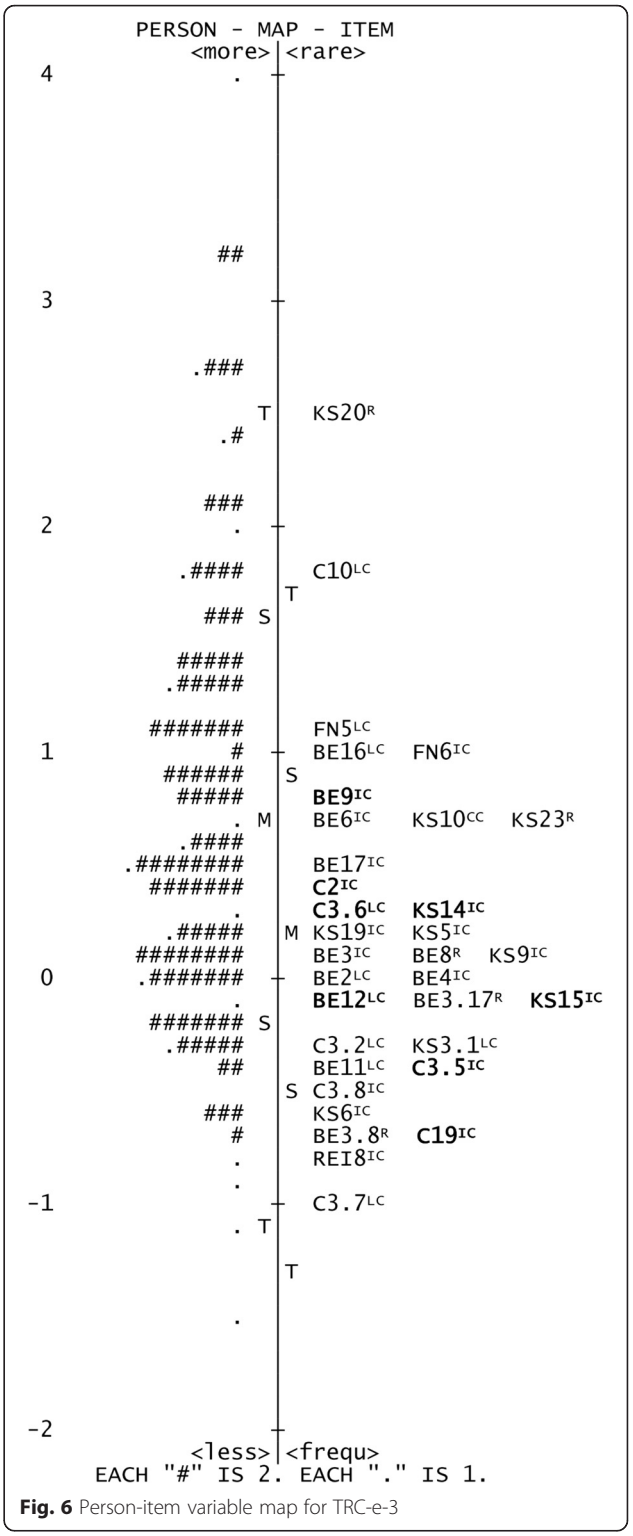

Fig. 6 Person-item variable map for TRC-e-3

0.65) for the TRC-e-4. None of the items presented infit or outfit values greater than 1.5. Regarding person fit statistics for the final TRC-e test forms, one second grader (0.4 % of the second grade sample), eight third graders (3.6 % of the third grade sample) and seven fourth graders (2.6 % of the fourth grade sample) obtained outfit values that were greater than 1.5. Compared to the percentage of misfit students in the calibration of the initial pool of items, the percentage of misfit students decreased in the second and fourth grade groups and increased only slightly (one student) in the third grade after linking.

Items are identified by the text to which they are related (FN = Fireflies' Night; C = The Caravels; BE = Observing Birds in the Estuary; KS = The King D. Sebastian), followed by the item's number. The comprehension level assessed by each item is presented in superscript (LC = literal comprehension; IC = inferential comprehension; R = reorganization; CC = critical comprehension). Common items appear in bold (Figs. 2 and 3 – common items

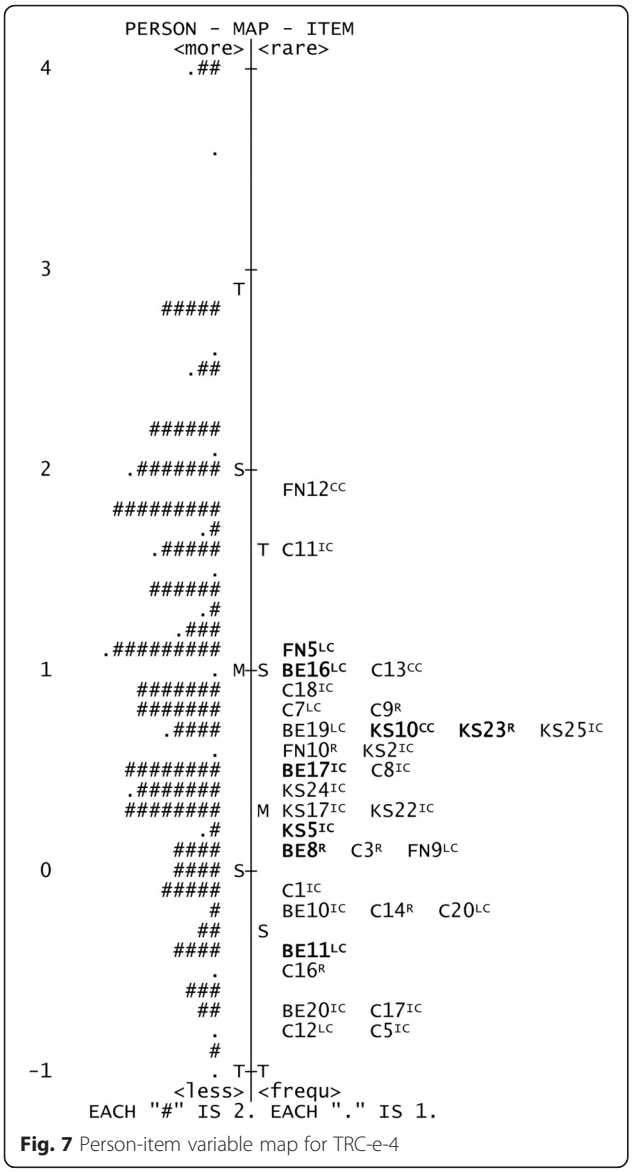

Fig. 7 Person-item variable map for TRC-e-4

between grades 2 and 3; Fig. 4 – common items between grades 3 and 4).

Item difficulty of the TRC-e test forms was now more appropriate for each target grade, given that none of the items was located outside of the range of each group's ability (−2.47 to 2.91 for the second graders, −1.46 to 3.96 for the third graders and −1.01 to 4.00 for the fourth graders). Mean values of the person ability scaled scores for the TRC-e were 0.13 (SD = .77) for the TRC-e-2, 0.70 (SD = .92) for the TRC-e-3 and 0.98 (SD = .98) for the TRC-e-4. As grades increased, person ability values were significantly greater, F (2, 724) = 42.060, $p < .001$. Bonferroni post-hoc tests revealed significant differences ($p < .05$) between the scaled scores obtained on the three TRC-e test forms.

Results of the PCA of the residuals of the TRC-e final test forms revealed that all the secondary dimensions had eigenvalues less than 2.0. The correlations of the test items' residuals ranged from zero to .24. In terms of the reliability of the TRC-e, the PSR coefficients were also moderate (TRC-e-2 = .72, TRC-e-3 = .77, TRC-e-4 = .78), the KR20 values were moderate to high (TRC-e-2 = .75, TRC-e-3 = .81, TRC-e-4 = .85) and the ISR coefficients were very high (TRC-e-2 = .95, TRC-e-3 = .95, TRC-e-4 = .95).

Discussion

This study examined the construction of two original reading comprehension tests for Portuguese students enrolled in second through fourth grade of the elementary school – the TRC-n and the TRC-e through the application of the Rasch model analyses. The development of these tests took into consideration theoretical and methodological guidelines concerning: (a) differences in performance when reading narrative or expository texts; (b) the demands of reading comprehension tasks with regard to the comprehension levels; and (c) the need of instruments that enable monitoring of inter- and intra-individual changes in reading comprehension over elementary school grades.

Test items were studied via Rasch model analyses. This allowed items to be selected for each test form according to the following criteria: (a) the difficulty of the items and persons' ability; (b) the results of infit and outfit mean square fit statistics; (c) point-measure correlation coefficients between each item and the total score; and (d) the mean ability value for persons who chose each option relative to each item.

For an accurate measurement, the item difficulty should match the persons' ability (Baghaei, 2008). For this reason, items that were extremely easy or extremely difficult for each grade were removed from the test form. The results of the fit statistics indicated inadequate outfit values for some items and some persons in both tests and in the three grades. The outfit statistic is sensitive to the presence of outliers, i.e., "unexpected behavior by persons on items far from the person measure level" (Linacre, 2011, p. 265). It implies that students with high ability level failed to answer to easy questions and low ability students succeeded on extremely difficult items (de Ayala, 2009). This misfit may be a consequence of ambiguous item wording, random answers or distraction, lack of cooperation and motivation; it is also possible that the misfit items were not working well with the bulk of the test items (Bond & Fox, 2007; Karabatsos, 2000; Linacre, 2011). Point-measure correlation coefficients were negative for some items; therefore, they were not properly measuring the construct. Following the guidelines by Wright and Linacre (1994) and by Bond and Fox (2007), in the TRC-n and the TRC-e, the items

with high outfit values and with negative point-measure correlations were removed from the test forms where they did not work well.

The analysis of the mean ability value for persons who chose each option relative to each item confirmed the presence of unexpected responses in some items, where lower ability persons succeeded in responding to higher item difficulty items and higher ability level persons failed to answer to lower item difficulty questions. These misfit responses are those which do not fit the Rasch model and that may be related with guessability or mistake-ability (Linacre, 2011). The items under this condition were also removed from the respective test forms. Anchor items were selected to incorporate into each test form of the TRC-n and the TRC-e, following the recommended selection guidelines (Dorans et al., 2011; Huynh & Meyer, 2010). These items comprised approximately 20 % of the total items across adjacent test forms and were distributed across all ability levels. Anchor items assessed the four reading comprehension levels and were invariant across adjacent test forms. After anchor items were selected, unique items for each test form were chosen by considering the spread of their difficulty in each grade, similar difficulty levels between items and the comprehension level each item assessed. Final test forms of the TRC-n were each composed of 27 items, six of which were anchor items between adjacent forms. Each of the final test forms of the TRC-e included 33 items, eight of which were anchor items.

After the selection of the items, the quality of the test forms of the TRC-n and the TRC-e was improved as the difficulty of the items became lined up with the persons and none of the items presented misfit values (Baghaei, 2008; Linacre, 2011). Recalibrations of the final test forms resulted in a pool of items with difficulty levels that were more appropriate for each grade. The difficulty of each TRC-n and TRC-e form increased with more advanced grades, as expected of an instrument that measures a construct that changes over the course of learning (Hardy et al., 2011; Kolen & Brennan, 2014). Bonferroni post-hoc test results revealed that each test form of the TRC-n and the TRC-e was able to adequately discriminate the performance of students enrolled in different school grades.

The analysis of the PCA of residuals was calculated for each form of the TRC-n and TRC-e. The results support the unidimensionality of each test form showing that the common variance among item responses is explained by a single latent trait. Correlations between the items' residuals confirm the items' local independence and corroborate the unidimensionality of the test forms. All test forms presented adequate reliability coefficients (Linacre, 2011).

Further studies should focus on the collection of criterion-related validity evidence as well as predictive validity evidence to support the utility of these instruments in educational contexts. Future research is also suggested to create new reading comprehension tests that allow for comparison of student's performance in narrative and expository texts, by using the same research sample so that narrative and expository test forms for each grade can be vertically scaled with each other. Future research should also focus on identifying which test (the TRC-n or the TRC-e) is the best predictor of performance in various academic subjects, such as the natural sciences, history and mathematics, as reading comprehension is a critical competency that at least partially explains difficulties in native language and academic subjects that require reading texts.

Conclusion

The goal of this study was to provide an application of the Rasch model analyses in the construction of two reading comprehension tests with vertically scaled test forms. Rasch model analyses were essential to create these reading comprehension measures that make it possible to assess large student samples and compare performances from second to fourth grade. Through the Rasch model analyses, in conjunction with the analysis of the items' content, it was possible to select to each test a set of items with good psychometric characteristics. The reduction of items in each test according to the Rasch model criteria enabled the development of reading comprehension tests that are time saving for teachers and researchers, without compromising the assessment of the latent trait.

The TRC-n and TRC-e present themselves as innovative reading comprehension measures, useful not only for longitudinal reading comprehension assessment throughout elementary school, but also for providing differentiated results regarding comprehension performance when narrative or expository texts are used, two aspects that are not often considered in test construction. In addition, the increasing extension of the texts included in the TRC-n and TRC-e makes them closer to the ones that students have to manage in school and other social contexts where text comprehension is mandatory.

Due to the adopted methodology and theoretical background in its construction, the TRC-n and the TRC-e test forms will have important implications for practice in the context of reading comprehension assessment. Its administration will give teachers, psychologists and researchers in educational contexts the possibility of longitudinally monitoring inter- and intra-individual changes in reading comprehension and

describing learning patterns or trajectories throughout the elementary school using percentile scores. In addition, these test forms will enable the assessment of the effects of intervention programs and help direct formal instruction to students' needs, and thus improve teacher effectiveness.

Appendix A

"Two twins almost alike" – text of the TRC-n for second graders

Rodrigo and Frederico are two twins almost alike. They are both skinny and tall like towers. Both have red hair, very curly and, since they found out they have myopia, both wear glasses.

The twins are always dressed alike. Every day they insist on getting to school with the same pants and sweater and even a pair of trainers with the same color. Only the most attentive are able to notice that Frederico's trainers are always a little more worn out and dirty than his brother's.

Occasionally, ***Rodrigo and Frederico try to fool their friends and trick them, pretending to be the other one. However, their friends are rarely fooled*** because, although they are just alike on the outside, the twins are very different in everything else: ***Rodrigo is shy*** and Frederico is talkative; ***one likes football, the other loves books;*** one has a great sense of humor, the other is a very serious person.

Note: The bold italic text parts are common to the third grade text.

"Two twins almost alike" – text of the TRC-n for third graders

Rodrigo and Frederico are two twins almost alike. They are both skinny, tall and a little clumsy, ***both have very curly red hair***. Since last year, after they went to the optician, ***they both wear glasses because they found out they have myopia***.

Both love to try to fool their friends and trick them, pretending to be the other one. However, their friends are rarely fooled because they know them like the back of their hands and they know that, for example, ***Rodrigo is shy*** and Frederico is more brazen, ***Rodrigo prefers to read books*** and play on his computer and ***Frederico is more into*** riding his bike or ***playing football***.

Although they dress alike and have the same hairstyle, the twins are rarely doing the same things. Except when they decide to play tricks, and on those days, it is possible to see Frederico reading a book and Rodrigo playing football!

However, as you can imagine, these switches are always noticed: Frederico is actually trying to read a book but, distracted, he forgets to turn the page and stays there, half an hour if need be, just looking at the same paragraph, probably thinking about the football game he is missing…

His friends ask him:

– Rodrigo, we are ten pages ahead… What are you thinking about?

And, in that moment, Rodrigo, who is not really Rodrigo but Frederico, still tries to disguise, but then he starts giggling…

His friends look at each other and say:

– Frederico, that's you! We thought so… We were already suspicious…

When Rodrigo plays the ball, he doesn't go unnoticed, trying to run, dribble, shoot … and almost always fails all these attempts. Moreover, Frederico is a great player, one of the school's best and it's hard to live up to. So after some time of awkwardness, his friends also ask him:

– Frederico, what's the matter with you today? Are you injured or what?

And then a more insightful friend adds:

– Or are you Rodrigo? It seems so…

Notes: The bold italic text parts are common to the second grade text.

Appendix B

Table 8 Reading comprehension items

School grade	Unique items	Anchor item
Second grade	1. Why are Frederico's trainers more worn out and dirty than his brother's? a. Maybe because he is less careful than his brother. b. Maybe because his trainers are of poorer quality. c. Maybe because he plays more football.	1. Of the two brothers, Rodrigo is the most … a. distracted. b. shy. c. fearful.
Third grade	2. Their friends almost always find out when they switch identities because… a. They cannot stand too long without saying who they are. b. When they talk they have a different voice. c. They cannot keep up the switch for too long.	

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

Authors' contributions
All authors have approved the manuscript and have contributed significantly for the paper, specifically: SS, Master: Made substantial contributions to conception and design, acquisition of data, and analysis and interpretation of data; was involved in drafting the manuscript and revising it critically for important intellectual content; gave final approval of the version to be published and; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. IC, Ph.D.: Made substantial contributions to conception and design, acquisition of data, and analysis and interpretation of data; was involved in drafting the manuscript and revising it critically for important intellectual content; gave final approval of the version to be published and; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. FLV, Ph.D.: Made substantial contributions to conception and design; was involved in drafting the manuscript and revising it critically for important intellectual content; gave final approval of the version to be published and; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. GP, Ph.D.: Made substantial contributions to design, analysis and interpretation of data; was involved in drafting the manuscript and revising it critically for important intellectual content; gave final approval of the version to be published and; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. SCS, Master: Made substantial contributions to acquisition of data; was involved in drafting the manuscript and revising it critically for important intellectual content; gave final approval of the version to be published and; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. AGS, Ph.D.: Made substantial contributions to conception and design of data; was involved in drafting the manuscript and revising it critically for important intellectual content; gave final approval of the version to be published and; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. IR, Ph.D.: Made substantial contributions to conception and design, acquisition of data, and analysis and interpretation of data; was involved in drafting the manuscript and revising it critically for important intellectual content; gave final approval of the version to be published and; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests
The authors declare that they have no competing interests.

Received: 8 April 2016 Accepted: 14 June 2016
Published online: 29 June 2016

References

Afflerbach, P. (2004). *High stakes testing and reading assessment*. Maryland: National Reading Conference Policy Brief. Retrieved from http://www.literacyresearchassociation.org/publications/HighStakesTestingandReadingAssessment.pdf.

Alderson, J. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Andreassen, R., & Braten, I. (2010). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading, 33*(3), 263–283. doi:10.1111/j.1467-9817.2009.01413.x.

Baghaei, P. (2008). Rasch model as a construct validation tool. *Rasch Measurement Transactions, 22*(1), 1145.

Barrett, T. C. (1976). Taxonomy of reading comprehension. In T. C. Barrett (Ed.), *Teaching reading in the middle class* (pp. 51–58). Boston, MA: Addison-Wesley.

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing, 26*(3), 349–379. doi:10.1007/s11145-012-9372-9.

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Taxonomy of educational objectives. Handbook 1: Cognitive domain. *Reading Psychology, 29*(2), 137–164. doi:10.1080/02702710801963951.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*, 253–269. doi:10.1002/sce.20106.

Boone, W. J., Townsend, J. S., & Staver, J. (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education, 95*, 258–280. doi:10.1002/sce.20413.

Buesco, H. C., Morais, J., Rocha, M. R., & Magalhães, V. F. (2015). *Metas curriculares de Português do ensino básico [Curricular benchmarks for Portuguese language in basic education]*. Lisboa: Ministério da Educação e Ciência. Retrieved from http://www.dge.mec.pt/sites/default/files/Basico/Metas/Portugues/pmcpeb_julho_2015.pdf.

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

Cain, K. (2010). *Reading development and difficulties*. Chichester, UK: BPS Blackwell.

Català, G., Català, M., Molina, E., & Monclús, R. (2001). *Evaluación de la comprensión lectora: Pruebas ACL (1°-6° de primária) [Reading comprehension assessment: ACL Tests (from 1st to 6th grade)]*. Barcelona: Editorial Graó.

Chen, R. S., & Vellutino, F. R. (1997). Prediction of reading ability: A cross-validation study of the simple view of reading. *Journal of Literacy Research, 29*(1), 1–24. doi:10.1080/10862969709547947.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Dorans, N. J., Moses, T. P., & Eignor, D. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21–42). New York: Springer.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*(3), 515–528. doi:10.1037/A0027182.

Goff, D., Pratt, C., & Ong, B. (2005). The relations between children's reading comprehension, working memory, language skills and components of reading decoding in a normal sample. *Reading and Writing, 18*(7–9), 583–616. doi:10.1007/s11145-004-7109-0.

Hardy, M. V., Young, M. J., Yi, Q., Sudweeks, R. R., & Bahr, D. L. (2011). Investigating the content and construct representation of a common-item design when creating a vertical scale. New Orleans, LA: National Council on Measurement in Education.

Hess, K. (2007). *Reading development and assessment of early literacy: A review of the literature*. Utah: Utah Department of Education.

Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation, 15*(2), 1–8.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*, 152–176.

Karlsen, B., & Gardner, E. (1996). *Stanford Diagnostic Reading Test* (4th ed.). San Antonio, TX: Harcourt Assessment.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kolen, M. J., & Brennan, R. L. (2010). *Test equating, scaling and linking* (2nd ed.). New York: Springer.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

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

Linacre, J. M. (2011). *A user's guide to WINSTEPS and MINISTEP: Rasch-model computer programs. Program manual 3.72.0.*. Chicago, IL: Winsteps.com.

Mullis, I., Martin, M., Foy, P., & Drucker, K. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: Boston College.

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods, 40*(4), 1001–1015. doi:10.3758/BRM.40.4.1001.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Oxford, UK: Blackwell Publishing Ltd.

RAND Reading Study Group. (2002). *Reading for understanding toward an R & D program in reading comprehension*. Santa Monica, CA: RAND corporation.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Sim-Sim, I., & Viana, F. L. (2007). *Para a avaliação do desempenho de leitura [Assessment of reading performance]*. Lisboa: Ministério da Educação - Gabinete de Estatística e Planeamento da Educação (GEPE).

Spinillo, A. G., Hodges, L. V. S. D., & Arruda, A. S. (2016). *Reflexões teórico-metodológicas acerca da pesquisa em compreensão de textos com crianças [Theoretical and methodological reflections about research on texts' comprehension with children]*. Psicologia: Teoria e Pesquisa.

Sweet, A. (2005). Assessment of reading comprehension: The RAND Reading Study Group vision. In S. Paris & S. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 3–12). Mahwah, NJ: Lawrence Erlbaum Associates.

Taylor, B. M., & Pearson, P. D. (2005). Using study groups and reading assessment data to improve reading instruction within a school. In S. G. Paris & S. Stahl (Eds.), *Childrens reading comprehension and assessment* (pp. 237–255). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2005). The CIERA school change framework: An evidence-based approach to professional development and school reading improvement. *Reading Research Quarterly, 40*(1), 40–69. doi:10.1598/RRQ.40.1.3.

Wilson, M., & Moore, S. (2011). Building out a measurement model to incorporate complexities of testing in the language domain. *Language Testing, 28*(4), 441–462. doi:10.1177/0265532210394142.

Woodcock, R., Mather, N., & Schrank, F. (2004). *Woodcock-Johnson III Diagnostic Reading Battery*. Rolling Meadows, IL: Riverside Publishing.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.