Psicothema

Psicothema 2013, Vol. 25, No. 3, 384-389
doi: 10.7334/psicothema2012.288

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Validity of a reading comprehension test for Portuguese students

Irene Cadime[1], Iolanda Ribeiro[1], Fernanda Leopoldina Viana[1], Sandra Santos[1], Gerardo Prieto[2] and José Maia[3]
[1] Universidade do Minho, [2] Universidad de Salamanca and [3] Universidade do Porto

Abstract

Background: The purpose of this work was to collect construct and criterion-related evidence of validity for a reading comprehension test (TCL - Teste de Compreensão da Leitura) with three vertically scaled forms, designed to assess students from second, third and fourth grade. **Method:** Two studies were conducted. In the first (n = 1,229), a confirmatory factor analysis was performed to analyse the test dimensionality. In the second (n= 402), concurrent and predictive evidence of validity was analysed using correlations between TCL, other reading tests and academic achievement. **Results:** Confirmatory factor analysis results supported a one-factor structure. Correlation coefficients with other reading tests were low to moderate and statistically significant. The TCL forms were shown to be good predictors of students' reading comprehension as assessed by teachers and of the National Exams of Portuguese Language results. **Conclusions:** Present results provide empirical evidence for the validity of the TCL forms.

Keywords: Reading comprehension, validity, reading assessment.

Resumen

Validez de un test de comprensión lectora para alumnos portugueses. **Antecedentes:** el objetivo de este trabajo fue recoger evidencia de validez de constructo y de criterio para un test de comprensión lectora (TCL- Teste de Compreensão da Leitura) con tres versiones escaladas verticalmente para evaluar alumnos portugueses de segundo, tercero y cuarto cursos de Primaria. **Método:** se efectuaron dos estudios. En el primero (n= 1,229) se analizó la dimensionalidad de la prueba recurriendo al análisis factorial confirmatorio. En el segundo (n= 402) se proporcionan datos sobre evidencia de validez concurrente y predictiva, analizando las correlaciones entre los resultados en TCL, los resultados en otras pruebas de lectura y los resultados académicos. **Resultados:** los análisis factoriales confirmatorios revelaron que el modelo de un factor se ajusta a los datos. Se obtuvieron coeficientes de correlación bajos a moderados y estadísticamente significativos entre las puntuaciones en el TCL y en otros tests de lectura. Las puntuaciones en TCL predijeron las puntuaciones obtenidas por los alumnos en los exámenes nacionales de lengua portuguesa y en las competencias de comprensión evaluadas por los maestros. **Conclusiones:** estos resultados proporcionan evidencia empírica para la validez de las versiones de TCL.

Palabras clave: comprensión lectora, validez, evaluación de la lectura.

Reading comprehension consists of the extraction and construction of meaning from interaction with a written text (RAND Reading Study Group, 2002). This definition entails two ideas: (a) readers play an active role in comprehending because they gather the meaning that the text explicitly conveys and construct their own meanings based on their background, and (b) at least two entities are involved in comprehension: the reader and the text. The simple view of reading (Hoover & Gough, 1990) proposes that reading comprehension is the product of accurate identification of the printed words (decoding) and the semantic and syntactic relationships among words and phrases (linguistic comprehension). In line with this view, significant correlations have been found between word decoding and recognition, linguistic comprehension and reading comprehension (Best, Floyd, & Mcnamara, 2008; Ouellette, 2006; Shankweiler et al., 1999).

Received: October 15, 2012 • Accepted: January 18, 2013
Corresponding author: Irene Cadime
Centro de Investigação em Psicologia
Universidade do Minho
4710 Braga (Portugal)
e-mail: irenecadime@psi.uminho.pt

Readers are expected to interact with the text to different degrees. Each task involved in comprehending a written text requires a different level of cognitive processing and demands the use of different sources of information (Basaraba, Yovanoff, Alonzo, & Tindal, 2013). Some tasks require only the detection and transcription of explicitly stated information in the text, while others demand the combined use of text information and previous knowledge of the reader. Several reading comprehension taxonomies have been formulated in an attempt to categorise the different demands of reading comprehension tasks (Barrett, 1976; Català, Català, Molina, & Monclús, 2001; Herber, 1978; Swaby, 1989). Although they were built with a number of different categories with distinct designations, some similarities can be found. Essentially, these taxonomies seem to converge in four domains designated by Català and colleagues (2001) as literal comprehension (LC), inferential comprehension (IC), reorganisation (R), and critical comprehension (CC). LC entails the recognition of information explicitly stated in the reading selection. IC emerges when the reader's prior knowledge is activated, and expectations and assumptions about the text contents are made based on clues provided by the reading. R implies a new way of organising information through synthesis,

schemes or summaries. CC includes making judgments with subjective answers, relating to the characters or the author's language and personal interpretations. This conceptualisation can allow teachers to design specific instructional activities to promote each type of comprehension. Tests should reflect those different types of comprehension in order to provide relevant feedback on students' achievement (Basaraba et al., 2013).

Formal measures of reading comprehension are scarce in Portugal. In 2007, a national study collected and evaluated the existing formal instruments of reading assessment (reading comprehension or decoding tests) validated for Portuguese students from first to sixth grades. Instruments were evaluated according to three global parameters: rationale (clarity of the objectives, theoretical basis), characteristics of the stimuli (adequacy of the texts and/or items), and psychometric characteristics (sample size and representativeness, items' difficulty and discrimination, reliability, validity, dimensionality). Results showed that there were no satisfactory reading comprehension instruments. The validation studies of the seven listed instruments were almost non-existent, and some tests also lacked explicit theoretical foundations (Sim-Sim & Viana, 2007).

In order to fill this gap, the construction of an original reading comprehension test – TCL-*Teste de Compreensão da Leitura* – was initiated in 2007, using the reading comprehension taxonomy of Català and colleagues (2001) as the theoretical guideline for item construction. This taxonomy was chosen because it synthesizes previous taxonomies and presents a clear operationalisation of each comprehension type. Three test forms were developed to assess students from second to fourth grade, using Rasch model analyses. The scores were placed on the same metric scale, through a vertical scaling process, so that the results obtained in different forms can be compared. The Rasch model, as an Item Response Theory model, overcomes several limitations of Classical Test Theory (for a review, see Hambleton & Jones, 1993) and is adequate to construct a common vertical scale for tests with distinct content to evaluate children with different levels of ability.

In 2008, a Spanish reading comprehension test (ACL), originally developed by Català and colleagues (2001), was adapted to Portuguese students from first to fourth grades, in the context of a master thesis (Mendonça, 2008). ACL and TCL share the same taxonomy as theoretical basis, the multiple-choice item format and include several types of text. However, ACL is composed by a distinct test form to each grade whose scores are not vertically scaled, thus not allowing the comparison of the gains obtained in reading comprehension from one grade to another. This is a fundamental limitation that TCL overcomes. Results of the adaptation study of the ACL test also presented poor criterion-related evidence of validity.

The main aim of the present investigation was to collect empirical evidence for TCL validity, using a two-step approach.

The goal of Step 1 was to analyse the dimensionality structure of the three TCL forms using confirmatory factorial analysis. The definition of reading comprehension – the extraction and construction of meaning from interaction with a written text – entails a unitary conception of the phenomenon. Although reading comprehension can be conceptualised using the taxonomies' types of comprehension, these operationalise the concept according to the tasks' demands and do not necessarily translate a multidimensional structure (Basaraba et al., 2013). This conception has been applied in the construction of reading comprehension assessments. The most widely used reading comprehension tests in English language (e.g., the reading comprehension subtests of the Gates-MacGinitie Reading Test [GMRT], the Nelson-Denny Reading Ability Test and the Scholastic Aptitude Test [SAT]) have a similar structure: a variety of texts that should be read silently, followed by multiple-choice questions that tap a range of abilities such as recalling specific information, making inferences, identifying the main idea or detecting the authors' tone. These abilities, which have a marked correspondence with the taxonomies' comprehension types, are assumed to be part of one single dimension called reading comprehension (Cook, Eignor, Steinberg, Sawaki, & Cline, 2009; Ozuru, Rowe, O'Reilly, & McNamara, 2008). Empirical evidence for a one-dimensional structure, using confirmatory analysis, has been provided for the GMRT (Cook et al., 2009) and for the SAT subtests (Dorans & Lawrence, 1999). Given the theoretical and structural similarities between TCL and the referred tests, a single dimension is expected to fit the data.

Step 2, the main goal of which was to collect criterion-related evidence, was divided into two situations: (a) results of the three TCL forms were compared to scores obtained in a word recognition test and in another reading comprehension test (concurrent validity); and (b) results of the TCL test forms were tested as predictors of teachers' evaluation of the students' reading comprehension and of the students' results in the National Exams of Portuguese Language (NEPL) (predictive validity). Reading comprehension is systematically assessed by teachers during classes, using informal tests constituted by texts from different types, followed by questions with varied formats (multiple-choice, open response, cloze, true/false). NEPL results are other indicator of comprehension achievement. Portuguese students take NEPL at the end of fourth and sixth grades. Only the first part of NEPL, which assesses reading comprehension, was considered. Teachers' evaluations and NEPL results are then external criterions that should be related to TCL's results.

Method

Participants

Data were collected from two different samples. For the study of dimensionality, a sample of 1,229 students from the second (n= 371), third (n = 403) and fourth grade (n = 455) was selected. Most of the participants (92.3%) attended public schools, with a few attending private schools (7.7%). Regarding gender, 60.6% of the students in the second-grade group, 50.4% of the third-grade group and 50.5% of the fourth-grade group were boys.

For the study of criterion-related evidence, a sample of 402 elementary education students was used. All attended public Portuguese schools. The students had the following grade distribution: 135 (33.6%) were second-grade students, 105 (26.1%) were third-grade students and 162 (40.3%) were fourth-grade students. Regarding gender, 64.4% of the students in the second-grade group, 47.6% of the third-grade group and 50% of the fourth-grade group were boys.

In both studies, all the participants were of Portuguese nationality and none had permanent special education needs.

Instruments

Reading Comprehension Test (TCL). TCL includes three forms – TCL-2, TCL-3 and TCL-4 – designed to assess the reading

comprehension skills of second, third and fourth-grade students. The same text is used in the three test forms. It integrates narrative, informative and instructional sequences, as well as poems. Items are multiple-choice with four options (one correct) and evaluate LC, IC, R and CC. Each test form has 30 items, being 30% anchor items. The Person Separation Reliability (PSR) and Item Separation Reliability (ISR) coefficients were high for TCL-2 (PSR = .70; ISR= .97), TCL-3 (PSR = .78; ISR = .98) and TCL-4 (PSR = .79; ISR = .98).

Word Recognition Test (*PRP - Prova de Reconhecimento de Palavras,* Viana & Ribeiro, 2010). PRP is comprised of 3 training items and 40 experimental items. Each item is composed of one image and four stimuli words, out of which only one corresponds to the image. Students must observe each image and choose the corresponding word by flagging it. PRP has a time limit of two minutes for third and fourth grade, and four minutes for first and second grade. It can be administered individually or in groups. Cronbach alphas ranged between .96 and .98. Test-retest reliability coefficients ranged between .76 and .88. Correlations with external criteria ranged between .36 and .62.

ACL Assessment of Reading Comprehension - forms ACL-2, ACL-3 and ACL-4 (Català et al., 2001; Portuguese adaptation by Mendonça, 2008). ACL-2, ACL-3 and ACL-4 were designed to assess students from the second, third and fourth grades. They include narrative and expositive texts, poems and graphical material. Items evaluate LC, IC, R and CC. ACL-2 is comprised of 24 items, ACL-3 of 25 items and ACL-4 of 28 items. Items are multiple-choice with four options for ACL-2 or five options for ACL-3 and ACL-4 (one correct). ACL can be administered individually or in group, without time limits. In the Portuguese adaptation, Cronbach alphas ranged between .78 and .83. In terms of convergent validity, a correlation of .33 was obtained between the results in ACL-2 and the results in a school's Portuguese language test, and a correlation of .20 was observed between ACL-2 and the classification of students' reading competencies performed by teachers. ACL-3 results had a correlation of .41 with the results in a school's Portuguese language test and a correlation of .36 with the classification of students' reading competencies performed by teachers. ACL-4 results had a correlation of .58 with the results of the NEPL.

Teachers' evaluation of students' reading comprehension. Results were collected from teachers in the end of the school year. They are expressed in a scale ranging from 1 (*poor*) to 5 (*excellent*) and are based on students' results in informal tests elaborated and scored by teachers.

National Exams of Portuguese Language (NEPL). The results from the NEPL, performed by students at the end of fourth grade, were collected. NEPL are conducted annually by the Portuguese Ministry of Education and the results are expressed in an ordinal scale ranging from A (*excellent*) to E (*poor*). The fourth-grade sample of the present study performed the NEPL of 2010, the third-grade sample performed the NEPL of 2011 and the second-grade sample performed the NEPL of 2012.

Procedure

Legal authorisations for data collection were obtained from the Portuguese Ministry of Education, school boards and parents. Trained psychologists administered tests over a two-month period during classes. All tests administrations followed the procedures indicated in their respective manuals, and the order in which the tests were administered was the same for all cohorts. The results achieved by students in the NEPL were collected from school boards in the years they were performed.

Data analyses

To empirically test the one-dimensional structure and the factorial validity of the three TCL forms, a confirmatory factor analysis (CFA) for each form's results was conducted using M*plus* software version 6.1 (Muthén & Muthén, 2010). The WLSMV estimator was used because of its robustness in dealing with categorical data, with samples higher than 200 and with a large number of variables (Muthén, Du Toit, & Spisic, 1997). Muthén has conducted unpublished simulation studies and found that sample sizes of 150 to 200 may be sufficient to medium-sized models—from 10 to 15 indicators (Brown, 2006). This has been confirmed by the simulation studies published by Flora and Curran (2004) under different sampling sizes (from 100 to 1000), varying degrees of non-normality and model complexity, which gives support to our approach given the present sample sizes and number of items per form.

Five criteria were used to evaluate the model's overall goodness of fit: (a) the Chi-Square Test of Model Fit (χ^2), (b) the Root Mean Square Error of Approximation (RMSEA), (c) the Comparative Fit Index (CFI), (d) the Tucker-Lewis Index (TLI), and (e) the Weighted Root Mean Square Residual (WRMR). The Chi-Square Test of Model Fit is an indicator of the discrepancy between the unrestricted sample covariance matrix and the restricted covariance matrix (Byrne, 2011). The higher the probability associated with the χ^2 value, the better the fit of the model. Therefore, p-values higher than .05 indicate a good model fit. RMSEA assesses the extent to which the co-variances implied by the parameters specified by the model correspond to the observed variances (Hoyle & Panter, 1995). RMSEA values of less than 0.05 indicate a good fit and values higher than 0.08 are not recommendable. CFI and TLI are related to the degree to which the model is superior to an alternative model that specifies the absence of co-variance among the variables in the reproduction of the co-variances observed (Hoyle & Panter, 1995). A CFI or TLI value higher than 0.90 is considered an indicator of good fit (Byrne, 2011). However, there are authors who suggest the adoption of a more restrictive criterion: a minimum value of 0.95 (Hu & Bentler, 1999). WRMR is based on the average difference between the observed correlation matrix and the matrix predicted by the model. Values lower than 1.00 can be considered indicators of good model fit (Yu, 2002).

Statistical analyses for the examination of criterion-related evidence of validity were performed with SPSS-*Statistical Package for Social Sciences,* version 17.0. To analyse the concurrent validity of the TCL forms, the results of TCL-2, TCL-3 and TCL-4 were correlated with the results obtained in PRP and ACL, as well as with the teachers' ratings and the results from the NEPL. Given that the distribution of the results in reading tests did not follow a normal distribution, and that teachers' ratings and the results of the NEPL are ordinal data, Spearman correlation coefficients were used. To analyse the predictive validity of the TCL forms, ordinal regression (SPSS Ordinal Regression procedure PLUM, with logit link function) was used, because both dependent variables were ordinal with five categories. After checking the proportional odds ratio assumption, six ordinal regression models were tested. In the first model, the results of TCL-2 were used as predictor variable

and the results in the NEPL of 2012 as the criterion variable. In the second, the results obtained by students in TCL-3 were used as the predictor variable and the results in the NEPL of 2011 as the criterion variable. In the third, the results of TCL-4 were tested as predictors for the results in the NEPL of 2010. In the three other models, the results of TCL-2, TCL-3 and TCL-4 were tested as predictors of teachers' evaluations.

Results

Construct-related evidence

The dimensionality of the TCL was studied by performing a CFA for each form to test a one-factor model.

Although the chi-square value is statistically significant for all three models, all other fit indices reached the values necessary to classify the models' goodness of fit as satisfactory (see Table 1). In all TCL forms, the RMSEA was less than .05 and the WRMR was less than 1.00. CFI and TLI values were high, being greater than .90 for the three test forms. However, if the more restricted criterion was adopted (Hu & Bentler, 1999), the CFI and TLI indices for TCL-2 did not reach the minimum value of .95 (see Table 1).

Criterion-related evidence

TCL-2, TCL-3 and TCL-4 results were significantly correlated with the results obtained in PRP, ACL, teachers' evaluation and results in the NEPL (see Table 2).

The correlations between the TCL results and other reading tests were positive, ranging between .23 and .73. Correlation coefficients between TCL and ACL were moderate for the third and fourth grades. For the second grade, the correlation was very low, contrarily to what was expected, given the theoretical and structural similarities of the instruments. The magnitude of the correlations between TCL and PRP was low. TCL-3 and TCL-4 correlations with PRP were lower than the ones observed with ACL (see Table 2).

Table 1
Global fit indicators

Form	χ^2 (405)	RMSEA	CFI	TLI	WRMR
TCL-2	471.69*	0.021	0.912	0.905	0.959
TCL-3	484.11*	0.022	0.953	0.949	0.947
TCL-4	473.05*	0.019	0.967	0.964	0.930

* p<.05

Table 2
Correlation matrix for TCL forms' results and other external criteria

	PRP	ACL	Teachers' evaluation	NEPL 2010	NEPL 2011	NEPL 2012
TCL-2	.39**	.23**	.55**			.47**
TCL-3	.36**	.66**	.65**		.60**	
TCL-4	.41**	.73**	.68**	.56**		

Note: NEPL= National Exam of Portuguese Language
** p<.01 (two-tailed)

The correlations between the TCL forms results, teachers' evaluation and the results of the NEPL were statistically significant, positive, and moderate, ranging from .47 to .60. The results of the ordinal regression models to predict the results on the NEPL revealed that scores obtained by students in each form of the TCL are statistically significant predictors (see Table 3). Results in TCL-2 accounted for 18% of the variance of results obtained in NEPL, while TCL-3 and TCL-4 accounted for 34 and 32% of the variance. The lower value observed for TCL-2, might be due to the time gap between the TCL's and the exam's administration.

The results of the ordinal regression models to predict teachers' evaluation of students' comprehension showed that scores in each form of TCL are statistically significant predictors of the teachers' evaluation (see Table 4). Results in TCL accounted for 27, 37 and 41% of the variance of results obtained in the teachers' ratings.

Discussion

The present investigation provides validity evidence for the three test forms that constitute the TCL. The results from the study of criterion-related evidence revealed a one-factor structure for the reading comprehension construct as measured by TCL-2, TCL-3 and TCL-4. This psychometric finding confirms that a total score for each test form can be used.

Table 3
Ordinal regression models predicting results in NEPL

	Beta	SE	Wald	OR	95% CI
Model 1					
TCL-2	0.24	0.06	19.22***	1.27	[1.14, 1.42]
Model 2					
TCL-3	0.32	0.06	27.96***	1.37	[1.22, 1.54]
Model 3					
TCL-4	0.28	0.04	48.70***	1.32	[1.22, 1.42]

Note: SE= standard error; OR= odds ratio; CI= confidence interval.
Model 1: $\chi^2_{(1)}$= 24.70, p<.001; Pseudo R^2= .18 (Cox & Snell). Model 2: $\chi^2_{(1)}$= 41.96, p<.001; Pseudo R^2= .34 (Cox & Snell). Model 3: $\chi^2_{(1)}$= 61.30, p<.001; Pseudo R^2= .32 (Cox & Snell).
*** p<.001

Table 4
Ordinal regression models predicting teachers' evaluation of reading comprehension

	Beta	SE	Wald	OR	95% CI
Model 1					
TCL-2	0.29	0.05	32.25***	1.34	[1.21, 1.48]
Model 2					
TCL-3	0.30	0.05	34.91***	1.34	[1.22, 1.48]
Model 3					
TCL-4	0.36	0.05	55.48***	1.43	[1.30, 1.57]

Note: SE= standard error; OR= odds ratio; CI= confidence interval.
Model 1: $\chi^2_{(1)}$= 43.27, p<.001; Pseudo R^2= .27 (Cox & Snell). Model 2: $\chi^2_{(5)}$= 48.61, p<.001; Pseudo R^2= .37 (Cox & Snell). Model 3: $\chi^2_{(5)}$= 86.71, p<.001; Pseudo R^2= .41 (Cox & Snell).
*** p<.001

The results of the study regarding concurrent and predictive evidence of validity, showed that the scores obtained in TCL-2, TCL-3 and TCL-4 are positively associated with scores on other reading tests and are good predictors of the results obtained in NEPL and in the evaluation of reading comprehension made by teachers. The magnitude of the correlation coefficients between TCL and PRP is lower than the ones observed in studies where the results in word recognition and reading comprehension tests were correlated (Best et al., 2008; Ouellette, 2006; Shankweiler et al., 1999). In these studies, correlations ranged between .47 and .89. The lower results obtained in our study may be due to the PRP format. Usually, word recognition is assessed by a task where the test takers must read aloud a list of words, without time limit. PRP requires the selection of the word that corresponds to each image, within a time limit. Lower correlations with PRP than with ACL were expected, given that PRP measures a related but not the same construct as TCL and ACL.

The validation study of ACL tests for Portuguese students provided correlation coefficients between teachers' evaluation of reading competences and ACL-2 and ACL-3, and a correlation between the results of a NEPL and ACL-4 (Mendonça, 2008). In the present study, the correlations obtained between the teachers' evaluation and the results in TCL are considerably higher than the ones obtained with the ACL tests. The correlations with the ACL tests were between .20 and .36, while correlations with TCL ranged from .55 to .68. The correlation coefficients with the results in NEPL are similar for both ACL and TCL. Notice that the correlations of ACL with external criteria are higher as the grade increases. A similar pattern was observed in our study: results in TCL-3 and TCL-4 had higher correlations with reading comprehension external criterions and explained more variance of the results in the regression models than TCL-2. It is possible that the performance of the second-grade students in reading comprehension tests is affected by the ongoing development of competences essential to comprehension, such as word recognition or reading fluency.

Limitations

The main limitation of this investigation is related to the use of the ACL test. In the study of adaptation for the Portuguese students, ACL tests presented good reliability indicators (Mendonça, 2008). However, concerning validity, the correlations between the results obtained by students in ACL-2 and ACL-3 and other external reading criteria were low. The structure of the ACL tests was not studied. Poor indicators of criterion validity and the lack of construct validity data for the ACL tests limit the inferences that can be made from the scores obtained. In the present study, it is possible that the low correlation obtained between TCL-2 and ACL-2 is due to the psychometric problems of the ACL test.

Guidelines for future research

Future research should include divergent validity studies of the TCL forms, by studying the association between the results obtained in the TCL and results on a test that measures a theoretically unrelated construct (e.g., numeric reasoning).

Additional evidence of predictive validity for the TCL forms should also be collected. In future studies, the results obtained by students in the sixth grade NEPL should be collected and used as a criterion to study the predictive properties of TCL.

The structure of the TCL forms should be further investigated using samples from other Portuguese speaking countries, so that cross-cultural comparisons can be made to study the factor structure stability.

Conclusions

The results indicate that TCL is a valid instrument for measuring the reading comprehension abilities of Portuguese students. Evidence of validity is an essential aspect to guarantee that a test adequately represents the construct that it intends to measure and that the proposed interpretations can be accurately made from the scores obtained (AERA, 1999).

The TCL was constructed as a response to the necessity of formal reading comprehension tests for Portuguese students in elementary school. The lack of formal tests with strong and well-studied psychometric properties limits the possibilities of identifying students performing below their reference grade group and examining performance changes. As a valid and reliable instrument, the TCL allows not only the detection of students who are performing poorly but also the evaluation of changes in reading comprehension across grades because total raw scores can be converted to percentiles and/or to standardised scores, which are placed on the same metric across test forms.

Acknowledgments

This research was supported by *Fundação para a Ciência e Tecnologia*, grant SFRH/BD/39980/2007.

References

American Educational Research Association (AERA) (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Barrett, T. (1976). Taxonomy of reading comprehension. In R. Smith & T. Barrett (Eds.), *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.

Best, R.M., Floyd, R.G., & Mcnamara, D.S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, *29*(2), 137-164.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Byrne, B. (2011). *Structural equation modeling with Mplus: Basic concepts, applications and programming*. New York: Routledge Academic.

Català, G., Català, M., Molina, E., & Monclús, R. (2001). *Evaluación de la comprensión lectora: Pruebas ACL [Assessment of reading comprehension: ACL tests]*. Barcelona: Editorial Graó.

Cook, L., Eignor, D., Steinberg, J., Sawaki, Y., & Cline, F. (2009). Using factor analysis to investigate the impact of accommodations on the scores of students with disabilities on a reading comprehension assessment. *Journal of Applied Testing Technology*, *10*(2), 1-33.

Dorans, N.J., & Lawrence, I.M. (1999). *The role of unity of analysis in dimensionality assessment*. Princeton, NJ: Educational Testing Service.

Flora, D., & Curran, P. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466-491.

Hambleton, R.K., & Jones, R.W. (1993). An NCME instructional module on comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Herber, H. (1978). *Teaching reading in content areas* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hoover, W.A., & Gough, P.B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127-160.

Hoyle, R.H., & Panter, A.T. (1995). Writing about structural equation models. In R.H. Hoyle (Ed.), *Structural equational modeling: Concepts, issues, and applications* (pp. 158-176). Thousand Oaks, CA: Sage Publications.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Mendonça, S. (2008). *Provas de avaliação da compreensão leitora: Estudo de validação [Reading comprehension tests: Validation studies]* (Unpublished master's thesis). University of Minho, Braga.

Muthén, B.O., Du Toit, S.H., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Technical report.

Muthén, B.O., & Muthén, L. (2010). *Mplus Version 6.1 [Software]*. Los Angeles, CA: Muthén&Muthén.

Ouellette, G. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*(3), 554-566.

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.

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

Shankweiler, D., Lundquist, E., Katz, L., Stuebing, K.K., Fletcher, J.M., Brady, S., Fowler, A., et al. (1999). Comprehension and decoding: Patterns of association in children with reading difficulties. *Scientific Studies of Reading*, *3*(1), 69-94.

Sim-Sim, I., & Viana, F.L. (2007). *Para a avaliação do desempenho de leitura [To evaluate reading achievement]*. Lisbon: Ministério da Educação - Gabinete de Estatística e Planeamento da Educação.

Swaby, B. (1989). *Diagnosis and correction of reading difficulties*. Boston: Allyn and Bacon.

Yu, C.Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Los Angeles: University of California.