EUROPEAN JOURNAL OF DEVELOPMENTAL PSYCHOLOGY, 2014
Vol. 11, No. 6, 754–766, http://dx.doi.org/10.1080/17405629.2014.938629

Routledge
Taylor & Francis Group

Developmetrics

Invariance on a reading comprehension test in European Portuguese: A differential item functioning analysis between students from rural and urban areas

Irene Cadime, Fernanda Leopoldina Viana, and Iolanda Ribeiro

University of Minho, Portugal

The aim of this study was to determine whether the items from a reading comprehension test in European Portuguese function differently across students from rural and urban areas, which biases the test validity and the equity in assessment. The sample was composed of 653 students from second, third and fourth grades. The presence of differential item functioning (DIF) was analysed using logistic regression and the Mantel–Haenszel procedure. Although 17 items were flagged with DIF, only five items showed non-negligible DIF in all effect-size measures. The evidence of invariance across students with rural or urban backgrounds for most of the items supports the validity of the test though the five identified items should be further investigated.

Keywords: Differential item functioning; Reading comprehension; Vocabulary; World knowledge.

Reading comprehension is the ability to extract and construct meaning from written language (Snow & Sweet, 2003). This construction of meaning involves different sources of information: to comprehend a text, readers rely on the explicitly stated information in the text and on their previous knowledge. Therefore, comprehension can be made at the literal level, or it can involve the elaboration of inferences (inferential comprehension), the synthesis or new ways

Correspondence should be addressed to Irene Cadime, Centro de Investigação em Estudos da Criança, Instituto de Educação, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal. E-mail: irenecadime@ie.uminho.pt

This research was supported by FCT (Fundação para a Ciência e a Tecnologia) [grant number FCOMP-01-0124-FEDER-010733] and the European Regional Development Fund (FEDER) through the European program COMPETE (Operational Programme for Competitiveness Factors) under the National Strategic Reference Framework (QREN) and by FCT [grant number SFRH/BD/39980/2007].

of organizing the information (reorganization) or the production of personal interpretations or judgments (critical comprehension).

Vocabulary and world knowledge are two of the main predictors of performance on reading comprehension tests (Best, Floyd, & Mcnamara, 2008; Ouellette, 2006). Therefore, the results of reading comprehension tests that use texts with vocabulary that is more frequent in some regions or areas can be biased by the provenance of the respondents. The same bias can affect the results of tests that require world knowledge that is specific to certain groups of respondents.

The TCL reading comprehension test (TCL-*Teste de Compreensão da Leitura*; Cadime et al., 2013) was developed to assess the reading comprehension of Portuguese students from the second to the fourth grades. It uses a text with a bucolic theme that includes vocabulary and describes situations that may be more familiar to students living in rural areas. For this reason, it is possible that some items may favour their performance on the test. This can threaten the test validity because the comparison of scores between students with the same level of the latent trait is biased by other students' characteristics. It is also essential that the measures guarantee the equity in testing, given that reading comprehension is often measured in research and educational settings to identify students performing poorly and therefore decide for their inclusion in special education and intervention programmes. If a large percentage of items favours one group over another, then the comparison of test scores between groups will be biased.

Differential item functioning (DIF) analyses are a micro-statistical procedure, performed at the item level, to assess the measure invariance across different groups of respondents (Walker, 2011).

The main goal of this study was to flag TCL items that may not be invariant across groups of students who live in Portuguese rural and urban areas and to assess the extent to which this lack of invariance might affect the validity of the test.

METHOD

Participants

The sample was composed of 211 second grade (mean age = 7.33, SD = 0.51), 196 third grade (mean age = 8.49, SD = 0.56) and 246 fourth grade students (mean age = 9.50, SD = 0.58). Data were collected in a rural area in the Minho region and in the metropolitan area of the city of Oporto in Portugal. Regarding the distribution by region, 64% of the second grade, 53.6% of the third grade and 65.9% of the fourth grade sample lived in the rural area. The two groups had similar age and reading level, and were similarly distributed by sex in third and fourth grades, but the second grade rural sample had a higher percentage of boys (see Table 1).

TABLE 1
Sample characteristics for each grade and provenance area ($N = 653$)

	Grade 2 ($n = 211$)				*Grade 3* ($n = 196$)				*Grade 4* ($n = 246$)			
	Rural		*Urban*		*Rural*		*Urban*		*Rural*		*Urban*	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Sex												
Female	48	35.6	38	50	55	52.4	38	41.8	81	50	42	50
Male	87	64.4	38	50	50	47.6	53	58.2	81	50	42	50
RL[a]												
>P50	108	80	65	86.7	63	60	65	73	116	71.6	55	67.1
Age												
M (SD)	7.30 (0.51)		7.38 (0.52)		8.41 (0.53)		8.58 (0.58)		9.56 (0.62)		9.38 (0.49)	
Range	7–9		7–9		8–10		8–10		9–12		9–10	

Note: RL, reading level given by the number of students in percentile 50 or above in word recognition proficiency as assessed by the test PRP (see the supplementary information).
[a] No information about RL was obtained for one second grade, two third grade and two fourth grade participants from the urban area.

Measure

The TCL reading comprehension test is composed of three specific test forms (TCL-2, TCL-3 and TCL-4) to assess second, third and fourth grade students. All three forms have a common text divided into sections with an extension between 41 and 372 words. Each test form is composed of 30 multiple-choice items with four options (one correct) that assess literal comprehension (12), inferential comprehension (9), reorganization (6) or critical comprehension (3). Confirmatory factor analysis provided evidence for a one-dimensional structure. Reliability coefficients varied between .70 and .98. See the supplementary information for more detailed information about the TCL's psychometric properties.

Procedure

Legal authorizations for data collection were obtained. Trained psychologists administered the test in groups, without time limits. Additional information about data collection is available in the supplementary information. Students with cognitive impairment who were eligible for Special Education were not included in the study.

Statistical analyses

DIF between students from rural and urban areas was investigated for each test form, taking students from the rural area as the focal group and using two procedures: logistic regression (LR) and Mantel–Haenszel (MH; Holland & Thayer, 1988).

LR was used to investigate the existence of uniform and non-uniform DIF. Two effect-size (ES) measures were obtained: (a) change in R^2 between the fitted models (ΔR^2) and (b) the delta log odds ratio (Δ_{LR}) proposed by Monahan, McHorney, Stump and Perkins (2007). The first was calculated for uniform and non-uniform DIF and the second only for uniform DIF. MH chi-square statistic with continuity correction and the ES Δ_{MH} were computed for each item to investigate the existence of uniform DIF. See the supplementary file for additional information about missing data and the LR and MH statistics.

The package difR (version 4.6) for R (Magis, Béland, Tuerlinckx, & De Boeck, 2010) was used to perform the analyses. Given that an item that presents differential functioning affects the validity of the total test scores (Navas-Ara & Gómez-Benito, 2002), an iterative purification process was used: items flagged as DIF are excluded from the total test score computation and the analysis is rerun; the process is repeated until two successive iterations return the same classification of the items as DIF or DIF-free (Magis et al., 2010).

Significance level was 5%. To evaluate the ΔR^2 magnitude, the standards proposed by Jodoin and Gierl (2001) were used: negligible if $\Delta R^2 < .035$; moderate if $.035 \leq \Delta R^2 < .070$ and large if $\Delta R^2 \geq .070$. An effect-size classification scheme based on the Educational Testing Service (ETS) classification was used to evaluate the Δ_{LR} and Δ_{MH} magnitude: negligible if $\Delta_{LR/MH} < |1|$; moderate if $|1| \leq \Delta_{LR/MH} < |1.5|$ and large if $\Delta_{LR/MH} \geq |1.5|$.

RESULTS

Students from the urban area obtained higher total scores than students from the rural area in second ($t_{(209)} = -5.57$, $p < .001$), third ($t_{(194)} = -3.91$, $p < .001$) and fourth ($t_{(244)} = -2.43$, $p < .05$) grades (see Table 2).

Items flagged as having some type of DIF by at least one procedure are presented in Table 3. Regarding the TCL-2, LR flagged item 17 as having uniform DIF and item 13 as having non-uniform DIF, but no item was flagged by MH. Item 17 tended to favour the focal group. On item 13, students from the rural area with low levels of reading comprehension tended to perform better than did students from the urban area (see Figure 1).

TABLE 2
Descriptive statistics for the TCL scores obtained by each group

	Total sample			*Rural*			*Urban*		
	Mean	*SD*	*Range*	*Mean*	*SD*	*Range*	*Mean*	*SD*	*Range*
TCL-2	11.97	4.82	1–29	10.67	4.36	1–23	14.28	4.77	5–29
TCL-3	15.63	4.89	4–26	14.40	5.08	4–25	17.04	4.27	8–26
TCL-4	16.95	5.05	5–27	16.39	4.98	6–26	18.02	5.05	5–27

Note: TCL-2 $N = 211$; TCL-3 $N = 196$; TCL-4 $N = 246$.

TABLE 3
Items flagged with DIF, types of DIF (uniform or non-uniform), test statistics and *p*-values for the two methods (LR and MH) and effect sizes (ΔR^2, Δ_{LR} and Δ_{MH})

		Logistic regression (LR)				*Mantel–Haenszel (MH)*		
Item	*Type*	*Statistic*	*p*	ΔR^2	Δ_{LR}	*Statistic*	*p*	Δ_{MH}
TCL-2								
13CC^{s7}	NUDIF	4.079	.043	.021 (A)				
17LC^{s9}	UDIF[a]	5.347	.021	.025 (A)	2.167 (C)	–	–	–
TCL-3								
6CC^{s5}	NUDIF	6.140	.013	.048 (B)				
1LC^{s1}	UDIF[b]	15.517	<.001	.093 (C)	3.036 (C)	10.903	.001	2.565 (C)
5IC^{s3}	UDIF[b]	4.786	.029	.031 (A)	2.062 (C)	4.357	.037	2.030 (C)
8R^{s7}	UDIF[a]	5.883	.015	.038 (B)	1.903 (C)	6.640	.010	1.836 (C)
9R^{s8}	UDIF[a]	3.893	.049	.024 (A)	1.825 (C)	–	–	–
10IC^{s9}	UDIF[a]	12.260	<.001	.078 (C)	3.100 (C)	11.587	<.001	2.949 (C)
16LC^{s11}	UDIF[a]	4.796	.029	.028 (A)	1.727 (C)	5.022	.025	1.928 (C)
28IC^{s21}	UDIF[a]	–	–	–	–	4.124	.042	1.917 (C)
TCL-4								
28CC^{s19}	NUDIF	5. 635	.018	.029 (A)				
17CC^{s13}	UDIF[a]	–	–	–	–	4.216	.040	1.709 (C)
19IC^{s14}	UDIF[b]	5.344	.021	.024 (A)	1.688 (C)	–	–	–
23LC^{s17}	UDIF[a]	–	–	–	–	4.167	.041	1.818 (C)
24LC^{s17}	UDIF[a]	4.648	.031	.022 (A)	1.707 (C)	6.668	.010	2.506 (C)
26LC^{s18}	UDIF[b]	4.377	.036	.022 (A)	1.828 (C)	–	–	–
29IC^{s20}	UDIF[a]	16.064	<.001	.085 (C)	3.338 (C)	16.631	<.001	3.871 (C)

Notes: DIF was established when at least one of the two test statistics (LR or MH) exceeded the 5% significance level (TCL-2 $N = 211$; TCL-3 $N = 196$; TCL-4 $N = 246$). Items are identified by the item number in the test form followed by the comprehension level assessed (LC, literal comprehension; IC, inferential comprehension; CC, critical comprehension; R, reorganization) and the text section that the item belongs to (in superscript); UDIF, uniform differential item functioning; NUDIF, non-uniform differential item functioning. The effect-size classification is indicated in parentheses: A—negligible ($\Delta R^2 < .035$ or $\Delta_{LR} < |1|$ or $\Delta_{MH} < |1|$); B—moderate ($.035 \leq \Delta R^2 < .070$ or $|1| \leq \Delta_{LR} < |1.5|$ or $|1| \leq \Delta_{MH} < |1.5|$); C—large ($\Delta R^2 \geq .070$ or $\Delta_{LR} \geq |1.5|$ or $\Delta_{MH} \geq |1.5|$).

[a] Item favouring the focal group, i.e. students from the rural area.

[b] Item favouring the reference group, i.e. students from the urban area.

Uniform DIF was detected in five TCL-3 items (1, 5, 8, 10 and 16) by both procedures. LR also flagged item 9 and MH flagged item 28 with uniform DIF. Most of these items tended to favour the rural area students. Students from the urban area tended to outperform the rural students only on items 1 and 5. Item 6 from the TCL-3 was flagged with non-uniform DIF: students from the focal group tended to perform better at the higher levels of the variable (see Figure 2).

Regarding the TCL-4, two items (24, 29) were flagged with uniform DIF by both procedures. Four items (17, 19, 23 and 26) were identified by only one procedure. Most of the TCL-4 items tended to favour students from the focal group. Non-uniform DIF was detected on item 28 from the TCL-4 (see Figure 3).

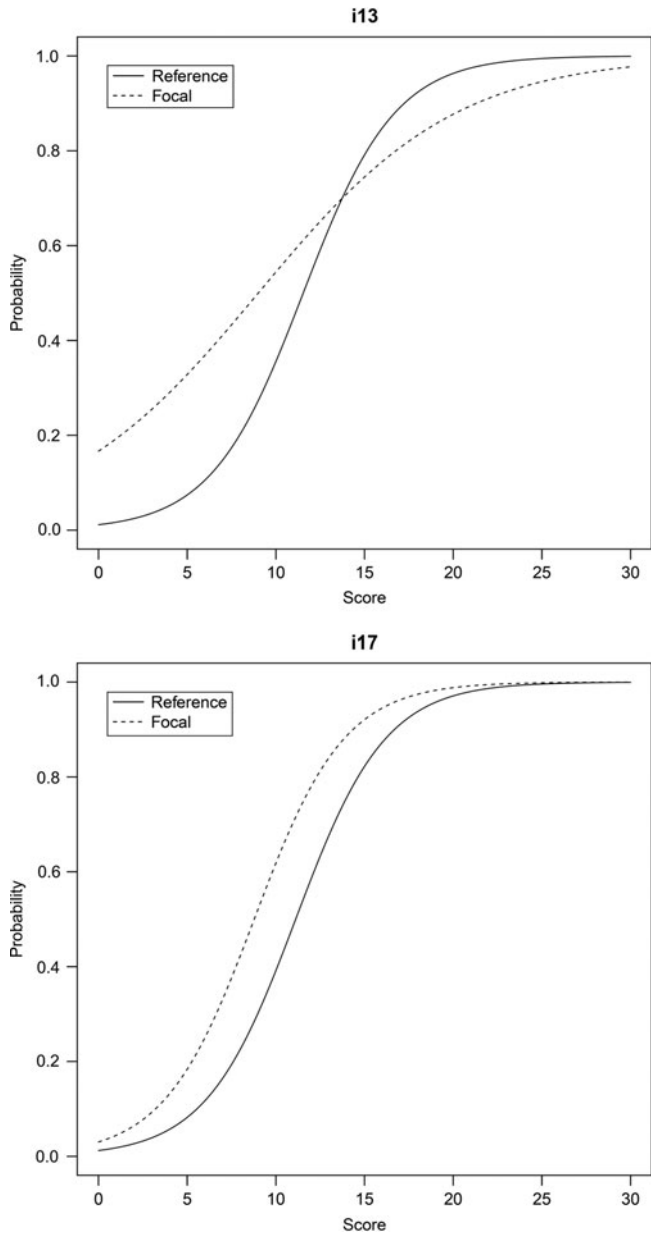

Figure 1. Logistic curves of the items of TCL-2 flagged as showing DIF by the logistic regression procedure.

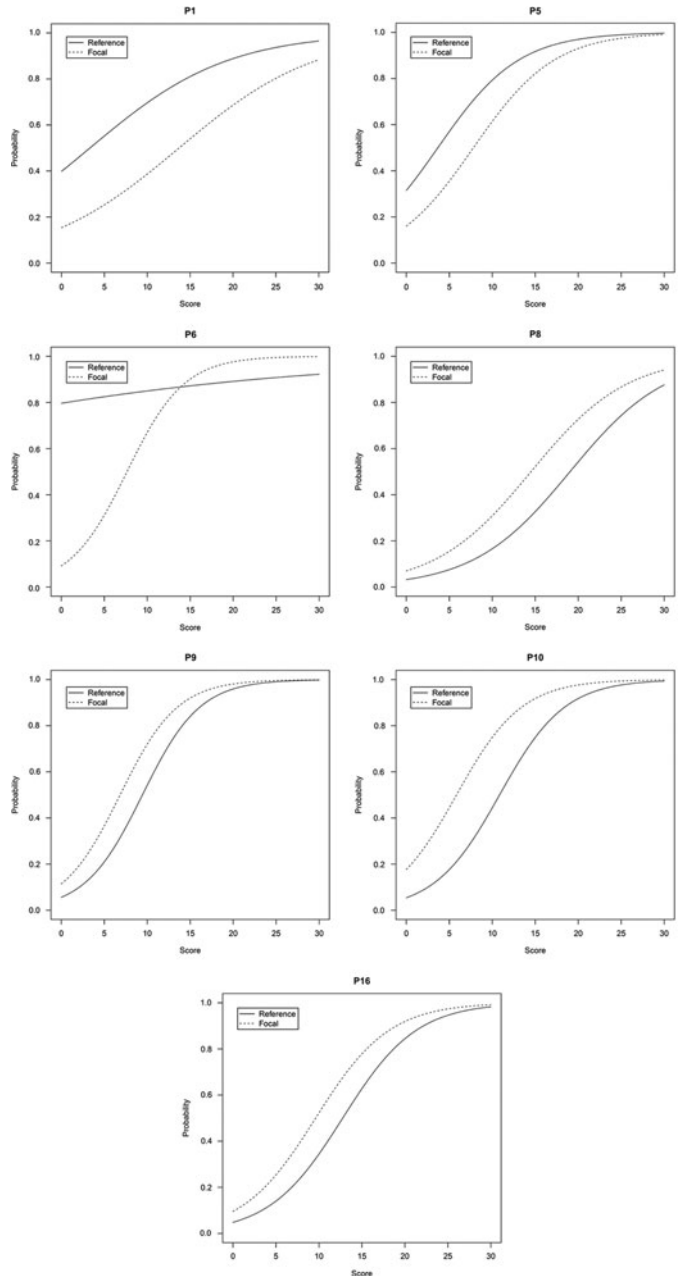

Figure 2. Logistic curves of the items of TCL-3 flagged as showing DIF by the logistic regression procedure.

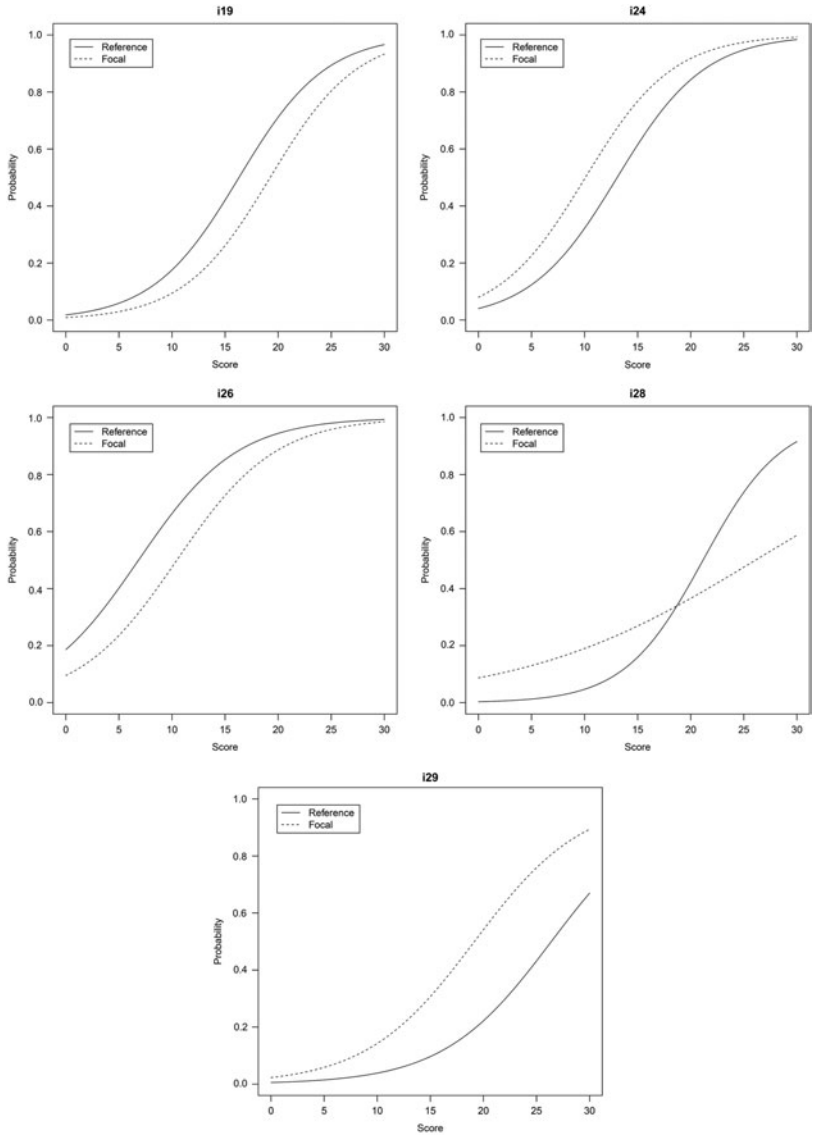

Figure 3. Logistic curves of the items of TCL-4 flagged as showing DIF by the logistic regression procedure.

According to the classification based on the ETS standards, all TCL items flagged with uniform DIF showed a large ES (see Table 1). However, when considering the ΔR^2 magnitude, most of these items can be classified as having

negligible DIF. Only items 1, 8 and 10 from TCL-3 and item 29 from TCL-4 showed a moderate/large DIF effect in all three ES measures. Regarding the items flagged with non-uniform DIF, only item 6 from TCL-3 had non-negligible DIF.

DISCUSSION

The main goal of this study was to detect items on which students from rural and urban areas would function differently on the three test forms of the TCL and to evaluate the extent to which these items threaten the test validity and introduce bias in the groups' comparison.

Seven items (TCL-3: 1, 5, 8, 10 and 16; TCL-4: 24 and 29) were flagged by both procedures as having uniform DIF. However, some discrepancies between the three ES were found. From these seven items, only four (TCL-3: 1, 8 and 10; TCL-4: 29) had non-negligible DIF in all ES measures. The remaining items had negligible DIF according to the Jodoin and Gierl's criteria for evaluating ΔR^2 magnitude, but large DIF according to the classification system for evaluating Δ_{LR} and Δ_{MH}. Similar results have been obtained in a study by Fidalgo, Alavi and Amirian (2014): compared with the ETS classification system, the Jodoin and Gierl's criteria also identified a lower number of items with non-negligible DIF. As Fidalgo, Alavi and Amirian (2014) point out, there are no simulation studies that determine the effectiveness of the various classification systems with small sample sizes and therefore further simulation studies are needed not only to determine their effectiveness but also to derive adequate cut-off points to be used with small samples. Therefore, items flagged in this study as having uniform DIF with large values for some ES and negligible values for other ES should be further investigated to evaluate whether its detection is correct.

Five items deserve particular attention: the four items flagged with non-negligible uniform DIF by the two procedures and all three ES (TCL-3: 1, 8 and 10; TCL-4: 29) and the item flagged with moderate non-uniform DIF by LR (TCL-3: 6). One main finding is that not all five items benefited the students from the rural area. In fact, on the first TCL-3 item, the students from the urban area had a higher probability of success on the item. This is a literal comprehension item that requires only the selection of the explicitly stated information in the text (see Appendix) and does not seem to require any specific vocabulary or world knowledge. In contrast, items 8 and 10 from TCL-3 and item 29 from the TCL-4 seemed to favour students from the rural area, as did item 6, which favoured this group's highest performers. These items assess different types of comprehension and also do not seem to require any particular knowledge of vocabulary or situations that might be more frequent in rural areas. Therefore, the five items require further investigation owing to the absence of a pattern in the differences by which they favoured one or the other of the two groups.

Four of these five items are from the TCL-3, representing 13% of the total test form. The fifth item integrates the TCL-4, corresponding to 3% of the test form. These percentages are low, considering that is common for 10–15% of the items in achievement tests to have DIF (Narayanan & Swaminathan, 1994).

The main limitation of this study is related to the fact that students were recruited only in two specific regions of the north of Portugal. Another limitation is the use of only one procedure (LR) to investigate the existence of non-uniform DIF. Future replication studies should use a larger sample consisting of students from varied rural and urban areas, and apply more than one procedure to detect non-uniform DIF.

The low percentage of flagged items with non-negligible DIF and the fact that most of these items favour students from the rural area, which is the group with lower total scores, lead us to conclude that the impact of these items on the scores comparability and on the test validity is low. Therefore, the TCL total scores can be compared between the two groups without significant concerns for DIF-related measurement bias.

Supplementary material

Supplementary material is available via the "Supplementary" tab on the article's online page (http://dx.doi.org/10.1080/17405629.2014.938629).

Manuscript received 19 February 2014
Revised manuscript accepted 19 June 2014
First published online 17 July 2014

REFERENCES

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*, 137–164. doi:10.1080/02702710801963951

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*, 384–389. doi:10.7334/psicothema2012.288

Fidalgo, A. M., Alavi, S. M., & Amirian, S. M. R. (2014). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing*. doi:10.1177/0265532214526748

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329–349. doi:10.1207/S15324818AME1404

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847–862. doi:10.3758/BRM.42.3.847

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics*, *32*, 92–109. doi:10.3102/1076998606298035

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel–Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*, 315–328. doi:10.1177/014662169401800403

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment*, *18*, 9–15. doi:10.1027//1015-5759.18.1.9

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*, 554–566. doi:10.1037/0022-0663.98.3.554

Snow, C. E., & Sweet, A. P. (2003). Reading for comprehension. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 1–11). New York, NY: The Guilford Press.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*, 364–376. doi:10.1177/0734282911406666

APPENDIX

TCL items that evidenced non-negligible DIF in all the effect-size measures

Text excerpts	*Item*
Portuguese version (original language)	
[. . .] Dentro de casa cheirava a cera fresca, a frutos maduros e aos ramos de alfazema que, no final da Primavera, a avó pendurava nas traves de castanho velho que atravessavam o teto da sala grande.	TCL-3—Item 1 O que pendurava a avó nas traves da sala grande? a) Cera fresca; b) Frutos maduros; c) Ramos de alfazema; d) Nada.
[. . .] E que jantar! Sopa que a avó nunca dispensa, arroz de tomate malandrinho a escorrer da travessa e bolinhos de bacalhau, doirados, perfumados de salsa e saborosos como só ela sabe fazer. No fim, um grande prato de barro coberto de nuvens, brancas e fofas [. . .] Que linda sobremesa!	TCL-3—Item 6 Como descreverias o jantar da Maria? a) Salgado mas delicioso; b) Poderia ter mais comida; c) Pouco apetitoso; d) Completo e saboroso.
[. . .] Trilho o Parque Natural em busca da cada vez mais rara águia-real. De repente, [. . .] três animais [...]. São 14h50. A cabra-montês está de volta a Portugal!O Gerês não me parece o mesmo! [. . .]	TCL-3—Item 8 Se tivesses de dar um título à pequena notícia que acabaste de ler, qual escolherias? a) Em busca da águia-real; b) A extinção da cabra selvagem; c) Visita ao Gerês; d) O regresso da cabra-montês.
[. . .] Andávamos nesta brincadeira, a comer amoras - “Vê lá, Maria, não te piques!” [. . .]	TCL-3—Item 10 Porque disse a avó “Vê lá, Maria, não te piques!”? a) Porque ela podia cair; b) Porque ela podia picar-se nas pedras bicudas; c) Porque ela podia picar-se nos tojos; d) Porque ela podia picar-se nas silvas.
[. . .] O girassol, Vira que vira, Como quem dança Um tango a solo; [. . .]	TCL-4—Item 29 Porque se diz que o girassol vira “como quem dança um tango a solo”? a) Porque o girassol dança quando há música no ar; b) Porque as pessoas dançam nos jardins ao lado do girassol; c) Porque o girassol vira como as pessoas que dançam; d) Porque o girassol vira com o vento.

(*continued*)

APPENDIX – *continued*

Text excerpts	*Item*
Translation from Portuguese to English [. . .] Inside the house, there was a smell of fresh wax, late fruits and lavender branches that, at the end of the spring, the grandmother hung in the ceiling of the dining room.	TCL-3—Item 1 What did the grandmother hang in the ceiling of the dining room? a) Fresh wax; b) Late fruits; c) Lavender branches; d) Nothing.
[. . .] What a dinner! Soup, which grandmother eats every day, tomato rice and golden codfish balls, perfumed with parsley and tasty, the way she always does. In the end, a big dish with white creamy clouds [. . .] What a beautiful dessert!	TCL-3—Item 6 How would you describe Maria's dinner? a) Salty but delicious; b) There was little food; c) Not very appealing; d) Complete and delicious.
[. . .] I was in the National Park looking for the rare golden eagle. Suddenly, [. . .] three animals [...]. It is 14h50. The wild goat is back to Portugal! The Gerês National Park looks different to me! [. . .]	TCL-3—Item 8 Which title would you choose for the journalist's report? a) Finding the golden eagle; b) The extinction of the wild goat; c) Visiting the Gerês National Park; d) The returning of the wild goat.
[. . .] We were playing, gathering blackberries -"Be careful Maria. Do not hurt yourself!" [. . .]	TCL-3—Item 10 Why did the grandmother say "Be careful Maria. Do not hurt yourself!"? a) Because she could fall; b) Because she could hurt herself on the pointy stones; c) Because she could hurt herself if she touched the gorse; d) Because she could hurt herself if she touched the bramble bushes.
[. . .] The sunflower, Moves and moves, Like someone who is dancing A tango all by himself; [. . .]	TCL-4—Item 29 Why does the author write that the sunflower moves "like someone who is dancing a tango all by himself"? a) Because the sunflower dances when there is music in the air; b) Because people dance near the sunflower in the garden; c) Because the sunflower turns like people when they are dancing; d) Because the sunflower turns with the wind.