Reading & Writing - Journal of the Literacy Association of South Africa
ISSN: (Online) 2308-1422, (Print) 2079-8245

Investigating measurement invariance in PIRLS 2021 across English and isiZulu language groups

Authors:
Sinethemba Mthimkhulu[1]
Karen Roux[1]
Maryke Mihai[1]

Affiliations:
[1]Department of Science, Mathematics and Technology Education, Faculty of Education, University of Pretoria, Pretoria, South Africa

Corresponding author:
Sinethemba Mthimkhulu,
u21727393@tuks.co.za

Dates:
Received: 30 Sept. 2023
Accepted: 23 Jan. 2024
Published: 20 Mar. 2024

How to cite this article:
Mthimkhulu, S., Roux, K. & Mihai, M., 2024, 'Investigating measurement invariance in PIRLS 2021 across English and isiZulu language groups', *Reading & Writing* 15(1), a455. https://doi.org/10.4102/rw.v15i1.455

Background: PIRLS 2021 results revealed that South African Grade 4 learners performed significantly lower compared to other countries in reading comprehension and that they did not reach the standardised international mean score of 500. It was also evident from the results that English learners performed relatively higher than isiZulu learners.

Objective: This study was initiated to investigate whether possible measurement invariance can help explain the difference in the scores of English and isiZulu learners. Measurement invariance is concerned with the metric and functional equivalence of the items in the two languages examined.

Method: A quantitative approach was utilised to reanalyse the PIRLS 2021 numerical data sets. Descriptive statistics were used to analyse the mean and raw scores. Rasch analysis was conducted to look for any Differential Item Function (DIF) in the items.

Results: Statistics revealed that there is a significant difference in the mean scores of the two languages. Analysis of raw scores provided evidence that some of the items lacked metric equivalence. Rasch analysis displayed that some of the items lacked functional equivalence.

Conclusions: The significant difference found in the scores implies that Grade 4 learners need more attention mastering the skill of reading literacy. The lack of metric and functional equivalence suggests that items in the passages need to be investigated and improved.

Contributions: This study has contributed to the International Large-Scale Assessment literature specifically relating to the equivalence of ILSA across languages. A significant gap in the learners' scores was identified and measurement invariance helped explain the gap in the learners' reading scores.

Keywords: differential item functioning; equivalence; measurement invariance; PIRLS 2021; reading literacy; validity.

Introduction

At Grade 4 level, reading literacy is a fundamental skill that empowers individuals to thrive intellectually, socially, and economically. As such, 'Progress in International Reading Literacy Study 2021' (PIRLS) is aimed to provide the participating countries (such as Russia, Egypt and Singapore) with comparative data on the reading trends of Grade 4 learners across the participating countries (Mullis & Martin 2021). The PIRLS assesses the reading abilities of Grade 4 learners to determine how well learners comprehend what they read. The reading abilities explored in this study are vital because the PIRLS 2021 results in South Africa show that some learners have lower reading abilities than the others; therefore, this study explores the factor that might possibly explain the difference in the reading abilities (achievement scores). The different participating countries have different education systems which implies that learners' reading proficiencies vary by country. For example, the Russian education system is different from the South African education system. Therefore, PIRLS developed international benchmarks to measure where the various countries are located (Mullis & Martin 2021) and to measure what the learner can do at each interval of the PIRLS scale (international benchmarks).

The PIRLS assessment instruments are made up of informational texts and narrative texts. It is crucial to note that PIRLS international versions of the instruments (achievement booklets and background questionnaires) are developed in English and allow the participating countries to translate the instruments into and adapt them to their national languages (Wry & Mullis 2023). In South Africa, that meant translating the PIRLS instruments into the remaining 10 official languages to be in accordance with the *Language in Education Policy (LiEP), 1997* (Department of Basic Education [DBE] 1997). The LiEP stipulates that South African learners must be taught in their

Note: Special Collection: Literacy in practice.

Home Language (HL) from Grade 1 to 3; therefore, the Language of Teaching and Learning (LoTL) must be the learners' HL (Department of Basic Education [DBE] 1997). In the South African study of PIRLS, this means that, for instance, learners who were taught in isiZulu in the Foundation Phase (Grade 1 to 3) complete the PIRLS assessment in isiZulu. It is crucial to note that the HL is a language chosen by the school to be regarded as the learners' HL (DBE 1997) and that the LoTL is a language that is chosen to be used as a medium of instruction. In some cases, the HL chosen at school is not the learners' true mother tongue or the language spoken at home.

After each PIRLS cycle, a summary report is released (DBE 2023; Howie et al. 2006, 2011, 2017). It has consistently been reported by the DBE (2023) and Howie et al. (2006, 2011, 2017) that South African Grade 4 learners are performing poorly, compared to other participating countries. Furthermore, there are large discrepancies in the South African mean scores compared by language. In PIRLS 2021, South African Grade 4 learners scored 288 points (SE = 4.4), which is significantly below the PIRLS international mean score of 500 (DBE 2023). When the PIRLS score is broken down by language, it shows that isiZulu (267 score points, SE = 6.5) learners performed lower than those who took the test in English (387 score points, SE = 14.5). None of the South African languages reached the low international benchmark of 400 score points (DBE 2023), implying that South African learners cannot read for meaning and are unable to locate and retrieve explicit information. It is noteworthy to mention that the data for the PIRLS 2021 cycle were collected during the global pandemic (coronavirus disease 2019 [COVID-19]).

Considering these results, the aim of this study is to shed light on the nature of isiZulu learners' reading literacy, as reported in the PIRLS 2021 data. Specifically, item equivalence is investigated in one selected PIRLS 2021 passage ('The amazing octopus') to ascertain whether possible item bias might have contributed to the differences in learner achievement in isiZulu, compared to English scores.

As such the following research questions were posed:

- How do the overall Grade 4 English and isiZulu learners' reading literacy achievement scores differ on PIRLS 2021?
- To what extent can the difference in achievement be explained by possible measurement invariance between English and isiZulu responses during the PIRLS 2021?
- To what degree are the items post-translation functionally equivalent?

Metric and functional equivalence were explored to establish equivalence in the translated texts and items (*cf.* Mthimkhulu 2023). It was necessary to investigate item difficulty (metric equivalence) post-translation to ascertain whether the level of difficulty was the same in the two groups. Functional equivalence (item behaviour) was examined to determine whether the items behaved or functioned similarly in both groups. Using Rasch Measurement Theory (RMT), Differential Item Function (DIF) was calculated to determine whether learners who completed 'The amazing octopus' passage and items in isiZulu were disadvantaged by possible measurement invariance. For this study, RMT allows the examination of items to ascertain their function in slightly different groups taking the same assessment.

Literature review

Translation complexities in Progress in International Reading Literacy Achievement booklets

Alharbi (2017) mentioned that when translating constructs or texts from their Source Language (SL) to a Target Language (TL), the construct measured in the SL should be the same as in the TL. The SL is the language in which the construct is initially designed, and the TL is the language the construct is translated into. During PIRLS translations of achievement booklets, some of the words proved challenging or were difficult to translate into South African indigenous languages to a point where the words had been included in English (Howie et al. 2017; Van Staden 2006). That raised the research questions which are addressed in this article because the scholars (Alharbi 2017; Mthimkhulu 2023) postulated after translation that the construct measured should be the same, and that the learners' achievement should depend on the learners' proficiency in the subject matter (Roux, Van Standen & Pretorius 2022). Translation of instruments (achievement booklets) into different languages may present a threat to the validity and equivalence of the test, possibly resulting in test or item bias (Pena 2007). Furthermore, the availability of up-to-date African terminology to accommodate the modern scientific and technological terms is also adding to the complexities accounted by PIRLS translators. Some scientific terms are not available in African languages which makes it difficult to translate them (Prah 2006).

Content and construct validity

Validity was reviewed in association with equivalence. Content validity was concerned with the fact that, if the English and isiZulu texts were equivalent, the English and isiZulu learners should arrive at the same meaning when reading the passages and responding to the items (Markus & Smith 2010; Oluwatayo 2012). In terms of construct validity, the concern was whether the PIRLS assessments measured the construct (reading comprehension) in the same way in English and isiZulu. If the items show DIF it may be concluded the measured constructs are no longer the same (Combrinck 2020).

Metric equivalence

Pena (2007) viewed metric equivalence as the difficulty of an item expressed in two or more languages. As the PIRLS achievement booklets required translations and adaptations, authors deemed it necessary to establish metric equivalence in the items of the selected passages ('The amazing octopus')

to examine whether item difficulty is the same across the two languages. Metric equivalence is a quantitative way of assessing cross-cultural equivalence in translated texts and an essential feature in examining construct validity (Kim, Han & Philips 2003). It was anticipated that the degree of metric equivalence in the items may perhaps help explain the difference in the scores. Thus, various item difficulties could be interpreted as a contributing factor to the difference in the scores. Cross-cultural equivalence has to do with the sameness of the instruments designed and translated into different cultures (Kim et al. 2003).

Functional equivalence

Functional equivalence is concerned with the instrument behaving the same way in different languages or cultural groups (Aegisdóttir, Gerstein & Cinnabars 2008). Pena (2007) mentioned that post-translation the instrument must elicit the same behaviour in the various languages the assessment is translated into. For this study, it meant that the instrument must function the same way in the English and isiZulu language groups.

Measurement invariance

Measurement invariance has to do with the psychometric equivalence of a construct across different groups taking the same instrument (Putnick & Bornstein 2016); however, if the construct has different meanings for the respondents, then it results in measurement non-invariance. If the instrument violates the condition of measurement invariance, the items of the instrument need to be examined (Combrinck 2020). Measurement invariance may be examined using different statistics such as individual item-fit statistics and Item Characteristics Curve graphs (ICC). It is postulated that the translation of instruments might have contributed to the measurement invariance, resulting in the two language groups experiencing challenges in understanding the texts and thus responding differently to the set of items in the passage. It is noteworthy that the difference in the achievement scores may be due to factors other than measurement invariance, factors such as teachers' qualifications, teachers' content knowledge, learners' attitudes towards reading and the availability of resources at home or at school.

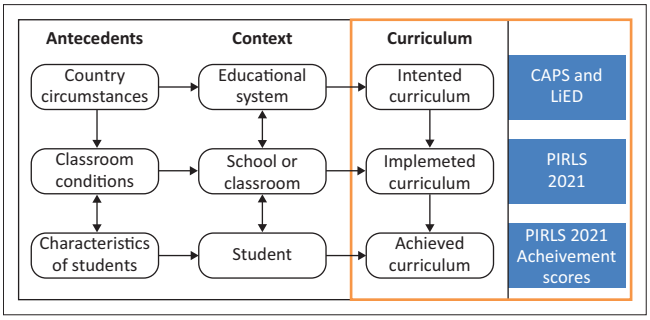
Theoretical framework

The curriculum process framework created by the International Association for the Evaluation of Educational Achievement (IEA) in 2005 (Mullis et al. 2007) was used as a theoretical framework for this investigation. The framework has three levels, namely the intended curriculum, the implemented curriculum, and the achieved curriculum (Organisation for Economic Co-operation and Development [OECD] & Programme for International Student Assessment 2006). The intended curriculum is produced by those in charge of the education system; the implemented curriculum is concerned with what happens in the classroom, while the achieved curriculum has to do with what learners achieve after their learning experiences. This framework is aimed at linking what is envisioned by the education system to what is implemented, subsequently evaluating the learning experiences of the learners through their performance assessment. Figure 1 depicts the curriculum process framework and the sections linked to this article.

For our purposes, the intended curriculum is linked to the Curriculum Assessment Policy Statement (CAPS) introduced by the (DBE 2011), and the LiEP introduced by the DBE (1997), as these documents articulate the envisioned education system in South Africa. The implemented curriculum is connected to PIRLS 2021 standardised assessments, meaning that PIRLS assessments are designed, taking into cognisance the different education systems of the participating countries. These assessments are in accordance with the LiEP (DBE 1997) because the IEA design PIRLS assessments and allow the participating countries to translate their assessments into their national languages. The LiEP states that the learners in Foundation Phase should be taught and assessed in their HL (DBE 1997). The achieved curriculum is linked to PIRLS 2021 achievement scores of the Grade 4 learners who completed the assessments in English and isiZulu. The rationale for the use of this framework is that it allows the investigation of the achieved curriculum, while considering the intended and the implemented curriculum. For this study, it accommodated the examination of the PIRLS achievement scores while considering CAPS and LiEP and PIRLS 2021 (cf. Figure 1).

Research methodology

The aim is to determine whether measurement invariance may have contributed to the difference in learner achievement scores of those who completed the PIRLS 2021 assessment in English and isiZulu. The quantitative approach was employed to draw data from the PIRLS 2021 data sets and, as such, this investigation becomes a secondary research study. Secondary data analysis is concerned with the reanalysis of the primary data to learn what remains to be learnt from the original data (Johnston 2014; McMillan & Schumacher 2014) and to enhance the primary research.



Source: Extracted from Organisation for Economic Co-operation and Development (OECD) & Programme for International Student Assessment, 2006, *Contextual framework for PISA 2006*, OECD, Paris

CAPS, Curriculum Assessment Policy Statement; LiEP, Language in Education Policy; PIRLS, Progress in International Reading Literacy.

FIGURE 1: Conceptual framework adapted from the curriculum process framework.

Participants

Since this study took the form of a secondary analysis of the PIRLS 2021 data, the South African Grade 4 learners, who completed the passage in English and isiZulu, were extracted from the original PIRLS 2021 Grade 4 sample. The PIRLS 2021 used Stratified Two-stage Cluster sampling design (Almaskut, LaRoche & Foy 2023) across all participating countries. The first stage of sampling meant sampling the schools, and the second stage of sampling meant sampling of classes within the school. The initial sample of the South African PIRLS 2021 main study comprised 12 426 Grade 4 learners stratified by language and province (DBE 2023). As the interest was in only two language groups, sample reduction took place. Furthermore, only one reading comprehension passage was utilised for the purpose of this investigation. Thus, only 505 Grade 4 learners were sampled based on the selected passage in the two languages. The sample of this study comprised English learners (n = 226) and isiZulu learners (n = 279) who responded to 'The amazing octopus' passage.

Data collection instruments

The PIRLS 2021 used two kinds of instruments, namely: the achievement booklets and background questionnaires. The former was made up of two types of texts: informational and narrative texts (Mullis & Martin 2021). It is the instrument that contains the passage and the items. It is noteworthy that 'The amazing octopus' passage was an informational passage. The latter instrument mainly focuses on gathering information about the social and educational environments of the learner (Howie et al. 2017). The PIRLS 2021 had validity and reliability procedures in place to ensure that comparative data were collected, including procedures such as quality assurance programmes (Johansone & Flicop 2023) and systems and instrument verifications (Ebbs et al. 2023; Von Davier et al. 2023). As such, this study adopted the validity and reliability procedures used by PIRLS 2021.

To fulfil the aim, only the data from the achievement booklets were utilised, with a specific focus on one passage ('The amazing octopus') and its 15 accompanying items. The items were both multiple-choice items and constructed responses. The 15 items in this passage were distributed in the PIRLS 2021 processes of comprehension (Mullis & Martin 2021). The passage examined was classified as an informational passage that focused on reading to acquire and use information. It is important to note that the data were collected during the global COVID-19 pandemic.

Data analysis

After data collection, the IEA Hamburg worked closely with the Trends in International Mathematics and Science Study (TIMSS) & PIRLS International Study Centre, Statistics Canada, and the National Research Coordinator (NRC) of the participating countries to engage in an extensive process ascertaining the integrity of the data. They are also responsible for the preparation of the international database to ascertain whether the data are valid, reliable, and comparable across countries taking part (Cockle 2023).

Since this investigation is about PIRLS 2021, it was necessary to make use of different statistical programmes that accommodated the complexity of the PIRLS data. These programmes included the International Database-Analyser (IDB-Analyser) plug-in with Statistical Package for Social Sciences (SPSS) (Foy 2018). The descriptive statistics of this inquiry included the calculation of mean scores and raw mean scores for the selected passage items in the two languages investigated.

After the initial analyses, DIF was calculated using RUMM2030. RUMM2030 is a statistical program that allows the user to put achievement data in and then it will run statistics and produce ICC graphs that determine the function of each item for the groups examined. Under the guidance of Rasch analysis, items were investigated to determine those that may have proved difficult and functioned differently in the two groups. The rationale for conducting DIF was to determine whether any of the items in the passage behaved differently and discriminated against any group sampled in this study. Boone (2016) and Combrinck (2018) postulated that DIF is a useful tool for investigating item bias against any other group who took the same assessment. Combrinck (2020) further mentioned that Rasch analysis places the respondents on the same scale as item difficulty and then detects where the item might have discriminated or functioned differently for persons in the lower and upper intervals across the investigated groups. Rasch analysis produced Analysis of Variance (ANOVA) statistics that enabled this study to compare the mean scores of the two groups, assign significance to the mean scores and to determine where the items might have functioned differently. The tables and figures presented below are data representations found upon the input and analysis of the PIRLS 2021 data.

Ethical considerations

An application for full ethical approval was made to the University of Pretoria, Faculty of Education, Ethics Committee and ethics consent was received on 30 November 2022. The ethics approval number is EDU161/22.

Results

Research question one focused on examining the mean scores and raw mean scores of those who completed the test in English and isiZulu. The overall mean score for South African Grade 4 learners is 288 points (SE = 4.4). Table 1 shows that a difference of 115 points exists between the mean scores of the two languages. To ascertain whether the

TABLE 1: Language comparison of the Grade 4 learners' achievement by language.

Language	Mean	SE	English	isiZulu
English	382	14.5	-	▲
isiZulu	267	6.5	▼	-

▲Significantly higher ▼ Significantly lower

difference in the means is significant, significance testing (t-test) was conducted.

Table 1 displays the significance of the two languages, mean scores. Upon testing, analysis revealed that there is a significant difference between the scores of those who completed 'The amazing octopus' passage in English and isiZulu. The English mean score is significantly higher than the isiZulu language mean score. It is crucial to note that the difference in the mean scores could be due to several factors, such as quality of translations, teachers' content knowledge, parental involvement, and teachers' qualification.

Table 2 displays the number and percentage of items that were correctly answered in the passage. They were calculated from the learners' raw scores. It is evident from Table 2 that learners who completed the passage in isiZulu found it very difficult, as none of the items were correctly answered by 50% of the learners. In contrast, 4 of the 15 items were correctly answered by 50% of the learners who responded to the passage in English. It is thus safe to say that 'The amazing octopus' passage lacked metric equivalence as evidence shows that for isiZulu language group the items were extremely difficult. Possibly, the different underlying abilities (learners' reading literacy skills) might factor in, regarding the level of item difficulty found by the learners in the distinct groups.

Following the results presented above, Rasch analysis was conducted to investigate whether any items in 'The amazing octopus' passage might have displayed DIF. Two of the research questions in this study mainly focused on determining whether item difficulty and item behaviour (possible item bias /measurement invariance) can help explain the significant difference found in the mean scores.

Table 3 presents the ANOVA statistics making it possible to determine which of the items in the passages discriminated between the two languages. Combrinck (2020:22) specified that a desirable chi-square is '$p > 0.05$' (as a probability smaller than 0.05 indicates acceptable model fit). The ideal range that indicates item discrimination among respondents is -2.5 to 2.5. If the residual value is positive that means that the item was too easy (underfit) for the respondents, whereas if the residual value is negative, it implies that the item was too difficult (overfit) for the respondents. Table 3 exhibits that 5 of the 15 items in the passage showed significant misfit where residuals were either above +2.5 or below -2.5. Items 7, 9 and 10 were underfit, inferring that they discriminated too little between the two language groups. Item 3 and 4 depicted overfit where the items were too discriminating within the two groups investigated. To further investigate the items, individual item-fit statistics for each language group was examined.

Table 4 indicates the individual item-fit statistics of the English language group only. It appears that none of the items in the passage show signs of misfit, either overfit or

TABLE 2: Number and percentage of items that were correctly answered in English and isiZulu.

Item No.	English			isiZulu		
	n Completed	*n* Correct	% Correct	*n* Completed	*n* Correct	% Correct
1	223	64	29	258	37	14
2	203	122	60*	232	65	28
3	178	70	39	221	57	26
4	190	94	49	211	81	37
5	196	37	19	230	4	2
6	203	97	48	226	10	4
7	202	122	60*	214	43	20
8	191	109	57*	185	38	21
9	208	49	24	198	3	2
10	185	67	36	173	3	2
11	191	83	43	173	37	21
12	185	51	28	168	5	3
13	170	85	50*	154	37	24
14	179	56	31	152	5	3
15	181	61	34	152	4	3

No., number.

*, Items correctly answered by 50% of the learners.

TABLE 3: Individual item-fit statistics for 'The amazing octopus' passage.

Item	Difficulty	SE	Fit residual	Chi-square	Probability
Z03	-0.260	0.133	2.651*	12.88	0.025
Z04	-0.895	0.124	2.855*	14.44	0.013
Z08	-0.727	0.130	1.688	13.84	0.017
Z13	-0.517	0.145	2.170	2.89	0.717
Z02	-1.090	0.120	-1.262	12.10	0.034
Z05	1.978	0.199	-1.441	4.46	0.485
Z07	-0.965	0.124	-3.253*	33.62	0.000**
Z09	1.678	0.187	-2.911*	14.54	0.013
Z10	0.922	0.168	-3.503*	19.59	0.001
Z11	-0.275	0.140	-1.622	8.64	0.124
Z15	0.899	0.173	-2.286	9.56	0.089
Z01	-0.363	0.081	1.839	12.72	0.026
Z06	-0.540	0.083	-2.458	11.58	0.041
Z12	0.265	0.107	-2.057	13.47	0.019
Z14	-0.110	0.083	-0.525	10.86	0.054

SE, standard error.

*, Fit residuals are shown when below -2.5 or above +2.5.

**, Bonferroni adjusted is 000667 for all the items. Items smaller than the Bonferroni adjustment are highlighted, and are significant.

TABLE 4: Individual item-fit statistics for English language group.

Item	Difficulty	SE	Fit residuals	Chi-square	Probability
Z03	0.272	0.199	1.444	9.64	0.008
Z04	-0.350	0.184	1.380	2.58	0.276
Z08	-0.822	0.183	1.723	13.75	0.001
Z13	-0.232	0.195	-0.158	4.31	0.116
Z02	-1.259	0.188	-0.594	1.57	0.456
Z05	1.881	0.226	-0.634	1.05	0.592
Z07	-1.284	0.187	-1.669	11.85	0.003
Z09	1.444	0.206	-1.979	8.94	0.011
Z10	0.537	0.194	-2.33	4.99	0.083
Z11	0.044	0.186	-1.141	4.27	0.118
Z15	0.679	0.199	-0.866	2.04	0.362
Z01	-0.047	0.119	1.757	3.84	0.147
Z06	-0.915	0.122	-1.42	3.06	0.217
Z12	0.123	0.134	-0.34	0.51	0.774
Z14	-0.072	0.103	0.406	3.71	0.157

SE, standard error.

*, Fit residuals are shown if below -2.5 or above +2.5.

**, Bonferroni adjustment is 000667 for all the items. All the items smaller than the Bonferroni adjustments are highlighted, and are significant.

underfit, within the English language group. It is evident from Table 3 that all the items for the English language group were a better fit.

Table 5 depicts individual item-fit statistics for the isiZulu language group. It is clear from the table that only one item depicted misfit. Item 11 showed underfit that was insignificant. The interpretation is that this item discriminated too little between the isiZulu learners who have lower and higher abilities. Upon determining that some of the items in the passage discriminated between the two language groups, DIF analysis was then conducted to investigate which of the items in the passage might have functioned differently for the two groups.

Table 6 presents evidence of the items that displayed DIF. Items 4, 7, 10 and 6 displayed signs of differential item functioning at a 5% significance level. It is evident that 4 of the 15 items present with significant uniform DIF. Uniform DIF implies that persons who have the same underlying abilities have consistently different probabilities of correctly responding to the item compared to the various groups (Andrich & Hagquist 2015). To further examine the data, ICC graphs were utilised to illustrate only the items that displayed differential item functioning from the Rasch analysis.

TABLE 5: Individual item-fit statistics for isiZulu language group.

Item	Difficulty	SE	Fit residuals	Chi-square	Probability
Z03	-0.919	0.167	1.875	1.837	0.399
Z04	-1.623	0.161	1.907	3.858	0.145
Z08	-0.759	0.193	0.822	5.014	0.081
Z13	-0.977	0.208	2.196	3.668	0.159
Z02	-1.066	0.163	-0.48	3.566	0.168
Z05	2.164	0.468	-0.776	0.507	0.776
Z07	-0.766	0.18	-1.623	6.956	0.030
Z09	2.586	0.628	-1.035	1.788	0.408
Z10	2.411	0.627	-0.908	1.773	0.412
Z11	-0.903	0.197	-2.864*	13.569	0.001
Z15	1.634	0.482	-1.011	0.555	0.757
Z01	-1.092	0.099	-2.366	5.169	0.075
Z06	-0.268	0.138	-1.064	2.554	0.278
Z12	0.126	0.197	-1.725	4.481	0.106
Z14	-0.547	0.133	-1.575	2.842	0.241

SE, standard error.

*, Fit residuals are shown if below -2.5 or above +2.5.

**, Bonferroni adjustment is 000667 for all the items. All the items smaller than the Bonferroni adjustment are highlighted, and are significant.

TABLE 6: Differential Item Function Summary of 'The amazing octopus' passage.

Item	F-ratio	Probability
Z08	7.81919	0.005463
Z04	13.71932	0.000243*
Z08	2.07984	0.150234
Z13	4.97365	0.026550
Z02	2.88598	0.090232
Z05	1.62755	0.202886
Z07	14.55166	0.000158*
Z09	8.66164	0.003469
Z10	24.20394	0.000000*
Z11	3.07292	0.080613
Z15	7.10563	0.008137
Z01	9.52798	0.002163
Z06	47.59605	0.000000*
Z12	10.28951	0.001491
Z14	0.20134	0.654006

F-ratio, Fisher's F ratio.

*, Significance at 5% level (Bonferroni 0.001111).

Item 4 (cf. Figure 2) in the passage investigated was a multiple-choice question expecting learners to focus on and retrieve Explicitly Stated Information. It asked whether octopuses have bones and what octopuses do. Among the different distractors, the learners were supposed to make one choice. For both groups at -2.5-person location, the sub-groups had the lowest probability of correctly responding to the item (< 30%). In isiZulu at -2 logits, they had approximately 50% chance of correctly responding to the item, compared to the English sub-group who had less than 10% chance at the lowest-class interval. At the same person location gradual increase in English is evident but the participants found the item to be difficult as their ICC is below the model curve. At the lower-class interval, the item was difficult for the English sub-group. For the isiZulu sub-group, at 0.5 logits, they had an similar probability of correctly responding to the item as the English sub-group at 1.5 logits (highest probability, approximately 80%).

Since item 4 was a multiple-choice item, Figure 3 illustrates a distractor analysis that provides different probabilities of learners choosing the correct answer or the distractors. The correct distractor for this item was C (3). At -3 to -2 logits, learners were tempted to opt for distractor B and distractor D (4). At that location learners had approximately a 20% chance of opting for distractor B. At 0 logits, learners had more than 50% chance of selecting distractor C (3) as the correct answer. Interestingly, at 1.5-person location learners had the highest probability of selecting the correct option (approximately 80% chance).

Item 7 (cf. Figure 4) also displayed DIF. It was a constructed response that required learners to make Straightforward Inferences. It asked what the octopus named Frieda learned to do. At -2.5 to -1.5 logits both sub-groups had the lowest probability of correctly responding to the question. At the lower-class interval, at -1.9, both sub-groups had less than 20% chance of getting the item correct. It is thus evident from

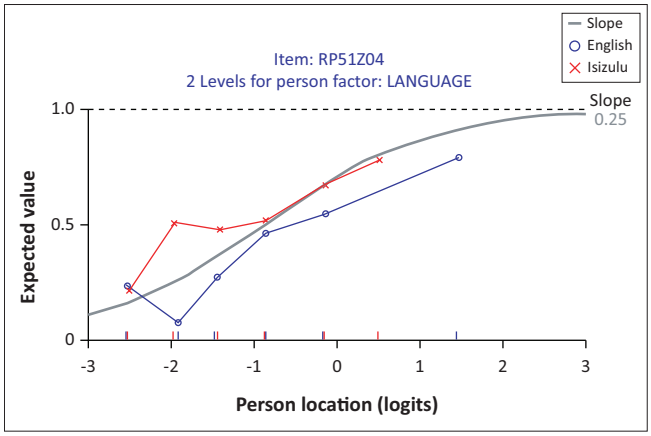


FIGURE 2: Item Characteristics Curve for item 4.

the graph that both groups experienced inconsistencies, while the item was difficult for both groups as neither of the ICC was above the model curve. However, from -1.5 logits, the English sub-group experienced an upward direction, and at -0.9 logits, they had approximately 50% chance of correctly responding to the item. As for the isiZulu sub-group, they also experienced a gradual increase in probability of getting the item correct at -2-person location. At the upper-class interval for the English sub-group, approximately 1.5 logits, they had the highest probability (95%) of responding correctly to the item. At the upper-class interval, specifically at 0.6 logits, the isiZulu sub-group had the highest probability of correctly responding to the item (100%). For both groups at the upper-class interval, they found the item not to be difficult as both their ICCs are above the model curve. Item 10 displayed DIF.

Item 10 (cf. Figure 5) also functioned differently in the two groups. It was a constructed response item that required learners to focus on and retrieve Explicitly Stated Information. The item was concerned with why the aquarium staff give octopuses puzzles. At the lower-class interval both groups had the lowest probability of correctly responding to the item. With the English sub-group at -2 to -0.7, they had the lowest probability of getting the item correct (approximately 10%). From -2.5 to -1.5 logits, isiZulu learners had the lowest probability of responding correctly to the item (0% chance). From -0.7 logits, the English sub-group experienced a steady increase in the ICC. In the isiZulu sub-group there was at 0.7 logits the lowest probability and approximately 30% chance of correctly responding to the item. At 1.8-person location, the English sub-group had the highest probability of correctly responding to the item (approximately 80% chance). At the lower-class interval both groups experienced difficulties in responding to the item, although at the upper-class interval the item was difficult and discriminating against the isiZulu sub-group. Next is item 6 in the passage that also displayed DIF.

Item 6 (cf. Figure 6) was the last of the 15 items of the passage that displayed DIF. It was a constructed response item that required learners to interpret and integrate Ideas and Information. This item required two constructed responses as it was a two-mark item. It required learners to provide two ways in which octopuses escape their predators. It is important to note that 1 mark is awarded if the learner provided a partially correct response (one response), and full marks are awarded for full comprehension when the learner provided two correct responses. For both sub-groups at the lower-class interval, specifically at -2.5 to -1.5, they had the lowest probability of correctly responding to the item (< 30% chance of scoring a partial mark). Interestingly, at that location (-1.5-person location), the English sub-group experienced an upward direction in their ICC, unlike the isiZulu group who had a gradual increase. At about -1 logits, the isiZulu sub-group

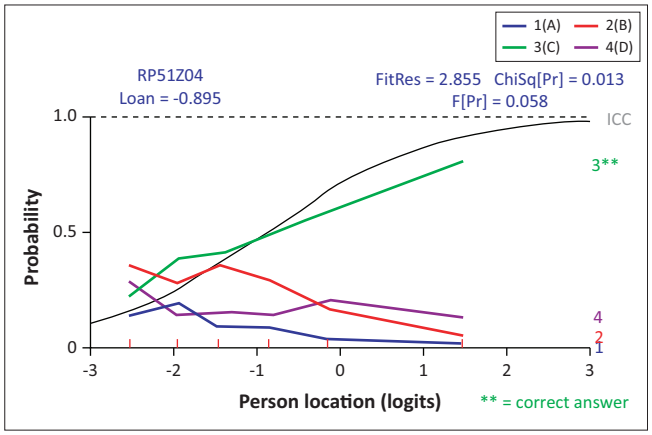


FIGURE 3: Distractor analysis for item 4.

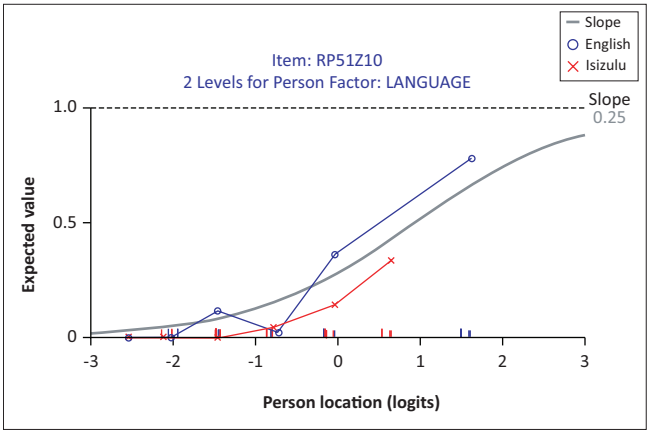


FIGURE 5: Item Characteristics Curve for item 10.

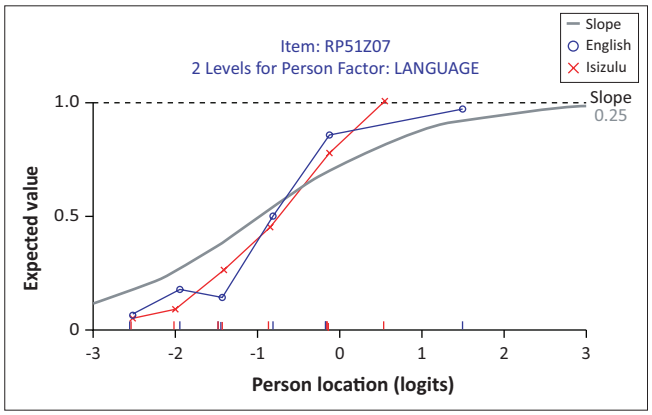


FIGURE 4: Item Characteristics Curve for item 7.

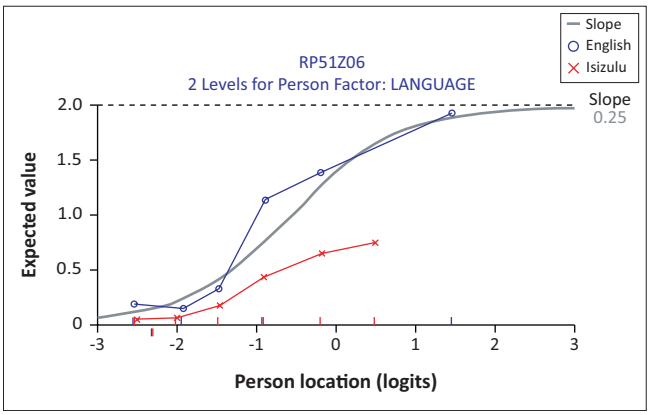


FIGURE 6: Item Characteristics Curve for item 6.

had the lowest probability (40% chance), whereas the English sub-group had the highest probability (100% chance) of correctly responding to the item. It is evidenced by the ICC that at the mentioned logits scale, the item was extremely difficult for isiZulu learners. In the upper-class interval, at 0.5-person location, isiZulu learners had the highest probability (approximately 70%) of correctly responding to the item and scoring a partial mark. As for English learners at the same person location, they had the highest probability of responding correctly to the item and obtaining full marks. It would appear that this item discriminated against the isiZulu sub-group in the upper-class as their ICC is below the model curve (inferring that they found the item difficult) and had the highest probability of scoring a partial mark in the item. Also, the item is difficult for the isiZulu sub-group as their ICC did not at any class move above the model curve.

Discussion of the results

Initially, this study stemmed from the poor literacy results found during PIRLS 2021 (DBE 2023). From the PIRLS 2021 South African report (DBE 2023), it was indicated that learners in South African are struggling to locate and retrieve explicitly stated information and that there is a discrepancy in the scores of those who tested in English and isiZulu. The significant difference found in the scores could be an indication that the isiZulu language learners need more attention (such as more reading material) when it comes to mastering the skill of reading literacy, compared to those who responded to the items in English. Table 2 presents evidence that the passage required a higher cognitive level of reading for isiZulu leaners. Although, it is also safe to infer that even for the English language group, the items were found to be difficult because not all the items were correctly answered by at least 50% of the learners. It is crucial to note that the implication is not that the passage was easy for English learners, but that isiZulu learners showed signs of struggling more with reading comprehension. When the raw mean scores were calculated they revealed that the isiZulu language had a mean score of 2 and the English group had 10. The 8-point difference in the raw mean scores may signify that for isiZulu learners the passage was demanding and thus that the learners need more reading materials to aid them in mastering the skill of reading literacy. With the items showing signs of metric inequivalence, it is postulated by Aegisdóttir et al. (2008) that the construct validity of the assessment is threatened.

It was suggested by Roux et al. (2022) that learners' achievement scores should reflect their reading proficiency, not anything else. Subsequently, item equivalence was then investigated to ascertain whether it is a contributing factor to the case of poor reading literacy achievement. Item-fit statistics showed that some items in the reading comprehension passage were misfit. The items that displayed misfits need to be paid attention to as some discriminated too much (overfit), and some discriminated too little (underfit), between the two groups. The implication is that these items did not have the same level of difficulty for both groups. It was stated by Combrinck (2020) that if the items show signs of DIF, investigations into the items should be done.

Further analysis was done to investigate item behaviour. It was found that some of the items in the passage behaved differently in the two groups. Items that displayed uniform DIF for this article meant that there is a lack of functional equivalence between the groups. The interpretation is that the difference in item behaviour might have contributed to different learner comprehension of the item, subsequently having different probabilities of responding to the item. Learners from the English and isiZulu groups, who have the same underlying abilities, did not have the same chance of responding correctly to the item across the two languages (Andrich & Hagquist 2015; Zumbo 1999). The presence of different item behaviour threatens content validity (Pena 2007) because the difference in the item behaviour implies that the learners might have experienced different item content or understood the item content differently.

Overall, the presence of lack of metric and functional equivalence in some of the items could explain the difference in the mean scores. In other words, it is found that the difference in the mean scores could be due to the violation of measurement invariance in the item of 'The amazing octopus' passage. However, it is critical to mention that the possible item bias found did not favour any language and the different ICC graphs showed that the four items functioned differently in each group; however, two of them were difficult for learners who responded to the items in isiZulu.

The evidence from this article suggests that language and literacy are interrelated, thus language plays a significant role in preparing learners to read to learn (reading literacy). It is for that reason that the DBE should carefully consider the effective implementation of the LiEP (DBE 1997), because scholars such as Coetzee-Van Rooy (2018), Nwammuo and Salawu (2018) and Nugraha (2019) stated that the policy is there but the implementation and the realisation of the LiEP (DBE 1997) is questionable. Therefore, a stronger emphasis should be placed on the realisation of language in education. The PIRLS 2021 indicated that 30% of the learners who took the PIRLS 2021 assessments indicated that the language of the test is not the language they authentically speak at home (Department of Basic Education [DBE] 2023). It, therefore, remains the responsibility of those in charge of the education system to ensure that there is effective implementation of the language policy. Nwammuo and Salawu (2018) made recommendations on how to raise awareness of and practise the language policy. They suggested that the curriculum must encourage the teaching of indigenous languages, with vital material resources; and by the updating and modernising of indigenous languages, the gap between policy and implantation might be bridged. Media can be utilised to

create public awareness and raise appreciation for the use of indigenous languages. If indigenous languages can be updated and modernised, translations may be improved and foreign words such as *i-Hammerhead shark* may be translated into the indigenous languages.

Recommendations and conclusion

Using the results and discussion sections the following recommendations are made:

- It is put forth that learners in South Africa should receive more challenging materials that are like the PIRLS achievement booklets, to get them familiar with the kind of texts PIRLS assesses them on and increase their vocabulary. In other words, a variety of texts for South African learners is recommended because the learners are usually exposed to literary texts. Exposure may also help develop learners' understanding of the world and its set of related systems while enriching their vocabulary.
- As the items in the passages display metric and functional inequivalence, it is recommended that these be refined and/or improved.
- Innovating methods of translation should be sought, and indigenous language speakers are tasked with the responsibility of updating and modernising the indigenous languages to fit into the current technological and scientific world.
- Lastly teachers may be provided with the necessary workshops and policy documents to help them effectively teach the reading literacy skill and help children master this skill.

In summary, reading literacy is a vital skill that every child enrolled in school should demonstrate. As the PIRLS data show that South African learners cannot read for meaning, it remains the responsibility of the nation to raise literacy levels and awareness. This study demonstrates that the isiZulu learners who completed the items in the examined passage found it difficult, compared to those who responded to the passage in English. The second analysis conducted in this investigation showed that some items in the passages functioned differently in the two languages investigated. This study provides evidence that the PIRLS passages examined lacked metric and functional equivalence to a certain degree as some of the items discriminated against and functioned differently in a certain group; subsequently they need to be improved wherever possible.

Acknowledgements

I thank LITASA for hosting a writing retreat and the participating professors for contributing valuable perspectives to this article.

This article is partially based on Mthimkhulu, S. dissertation entitled 'Examining PIRLS 2021: Differential item functioning across English and isiZulu language groups' towards the degree of Masters in Mathematics and Technology Education, University of Pretoria in 2023, with supervisor Dr Karen Roux and co-supervisor Dr Maryke Mihai. Dissertation not published at time of article publication.

Competing interests

The authors have declared that no competing interest exists.

Authors' contributions

K.R. assisted with data entry into RUMM2030, and M.M. served as an external advisor. S.M. composed the majority of the article, with assistance from K.R. and M.M.

Funding information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability

The new data that was created and analysed for this article is available upon reasonable request from the International Association for the Evaluation of Educational Achievement and Department of Basic Education.

Disclaimer

The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of any affiliated agency of the authors, or the publisher.

References

Aegisdóttir, S., Gerstein, H.L. & Cinnabars, D.C., 2008, 'Methodological issues in cross-cultural counselling research: Equivalence, bias, and translations', *Division of Counselling Psychology* 36(2), 188–200. https://doi.org/10.1177/0011000007305384

Alharbi, A.A., 2017, 'The translation and translation verification of the PIRLS Reading Questionnaires for Saudi students', Master's dissertation, Department of Educational Psychology, University of Kansas.

Almaskut, A., LaRoche, S. & Foy, P., 2023, 'Sample design in PIRLS 2021', in M. Von Davier, I.V.S. Mullis, B. Fishbein & P. Foy (eds.), *Methods and procedures: PIRLS 2021 technical report*, pp. 3.1–3.32, Boston College, TIMSS & PIRLS International Study Centre, Boston.

Andrich, D. & Hagquist, C., 2015, 'Real and artificial differential item functioning in polytomous items', *SAGE Journals* 75(2), 1–5. https://doi.org/10.1177/0013164414534258

Boone, J.W., 2016, 'Rasch analysis for instrument development: Why, when, and how?', *CBE Life Sciences Education* 15(4), 1–7. https://doi.org/10.1187/cbe.16-04-0148

Cockle, M., 2023, 'Creating the PIRLS 2021 international database', in M. Von Davier, I.V.S. Mullis, B. Fishbein & P. Foy (eds.), *Methods and procedures: PIRLS 2021 technical report*, pp. 7.1–7.16, Boston College, TIMSS & PIRLS International Study Centre, Boston.

Coetzee-Van Rooy, S., 2018, 'Dominant language constellations in multilingual repertoires: Implications for Language-in-Education Policy and Practices in South Africa', *Language Matters* 49(3), 19–20. https://doi.org/10.1080/10228195.2018.1493613

Combrinck, C., 2018, 'The use of Rasch measurement theory to address measurements and analysis challenges in social science research', PhD thesis, Department of Psychology, University of Pretoria.

Combrinck, C., 2020, 'Is this a useful instrument? An introduction to Rasch measurement models', in S. Kramer, S. Laher, A. Fynn & H.H. Janse van Vuuren (eds.), *Online readings in research methods*, pp. 11–41, Psychological Society of South Africa, Johannesburg.

Department of Basic Education (DBE), 1997, *Language in education policy*, Department of Basic Education, Pretoria.

Department of Basic Education (DBE), 2011, *Curriculum and Assessment Policy Statement (CAPS). Grades 4–6. English Home Language*, Department of Basic Education, Pretoria.

Department of Basic Education (DBE), 2023, *PIRLS 2021: South African preliminary highlights report*, Department of Basic Education, Pretoria.

Ebbs, D., Flicop, S., Hidalgo, M.M., & Netten, A., 2023, 'Systems and instrument verification in PIRLS 2021', in M. Von Davier, I.V.S. Mullis, B. Fishbein & P. Foy (ed.), *Methods and procedures: PIRLS 2021 technical report*, pp. 5.1–5.24, Boston College, TIMSS & PIRLS International Study Centre, Boston.

Foy, P., 2017, *PIRLS 2016 user guide for the international database*, TIMMS & PIRLS International Study Centre, Boston.

Howie, S.J., Combrinck, C., Roux, K., Tshele, M., Mokoena, G.M. & Macleod-Palane, N., 2011, *PIRLS Literacy 2016: South African children reading literacy achievement*, Centre for Evaluation and Assessment, Pretoria.

Howie, S.J., Venter, E., Van Staden, S., Zimmerman, L., Long, C., Du Toit. C. et al., 2006, *PIRLS 2006 summary report: South African children's reading literacy achievement*, Centre for Evaluation and Assessment, Pretoria.

Johansone, I. & Flicop, S., 2023, 'Quality assurance program for PIRLS 2021', in M. Von Davier, I.V.S. Mullis, B. Fishbein & P. Foy (eds.), *Methods and procedures: PIRLS 2021 technical report*, pp. 6.1–6.19, Boston College, TIMSS & PIRLS International Study Centre, Boston.

Johnston, M.P., 2014, 'Secondary data analysis: A method of which the time has come', *Qualitative and Quantitative Methods in Libraries 3*, 619–626, viewed 12 August 2023, from https://www.qqml-journal.net/index.php/qqml/article/view/169

Kim, M., Han, H. & Phillips, L., 2003, 'Metric equivalence assessments in cross-cultural research: Using an example of the Centre for Epidemiological Studies-Depression Scale', *Journal of Nursing Measurement* 11(1), 5–10. https://doi.org/10.1891/106137403780954930

Markus, K. & Smith, K., 2010, 'Content validity', in N Salkind (ed.), *Encyclopaedia of research design*, pp. 108–109, Sage, Thousand Oaks, CA.

McMillian, J.H. & Schumacher, S., 2014, *Research in education: Evidence-based inquiry*, 7th edn., Pearson Education, Boston, MA.

Mthimkhulu, S., 2023, 'Examining PIRLS 2021: Differential item functioning across English and isiZulu language groups', Master's dissertation, Department Science, Mathematics and Technology Education, University of Pretoria.

Mullis, I.V.S. & Martin, M.O., 2021, *PIRLS 2021 assessment framework*, Boston College, TIMMS & PIRLS International Study Centre, Boston.

Mullis, I.V.S., Martin, M.O., Kennedy, A. & Foy P., 2007, *PIRLS 2006 international report: IEA's Progress in international reading literacy study in primary schools in 40 countries*, Boston College, TIMSS & PIRLS International Study Center, Boston.

Nugraha, S.I., 2019, 'The language in education policy in South Africa: A gap between policy and efficacy', *Advances in Social Sciences, Education and Humanities Research 254*, 568–570. https://doi.org/10.2991/conaplin-18.2019.321

Nwammuo, A.N. & Salawu, A., 2018, 'Media roles in disseminating strategies for teaching and learning indigenous languages: The case of South Africa's language-in-education policy in post-apartheid era', *Cogent Arts & Humanities* 5(1), 1–13. https://doi.org/10.1080/23311983.2018.1553653

Oluwatayo, J., 2012, 'Validity and reliability issues in educational research', *Journal of Educational and Social Research* 2(2), 393. https://doi.org/10.5901/jesr.2012.v2n2.391

Organisation for Economic Co-operation and Development (OECD) & Programme for International Student Assessment, 2006, *Contextual framework for PISA 2006*. OECD, Paris.

Pena, D.E., 2007, 'Translation: Methodological considerations in cross-cultural research', *Child Development* 78(4), 1255–1261. https://doi.org/10.1111/j.1467-8624.2007.01064.x

Prah, K.K., 2006, *Challenges to the promotion of indigenous languages in South Africa*, Review Commission by the Foundation for Human Rights in South Africa, The Centre for Advanced Studies of African Society, Johannesburg.

Putnick, D.L. & Bornstein, M.H., 2016, 'Measurement invariance conventions and reporting: The state-of-the-art future directions for psychological research', *Developmental Review* 41, 71–79. https://doi.org/10.1016/j.dr.2016.06.004

Roux, K., Van Staden, S. & Pretorius, E.J., 2022, 'Investigating the differential item function of a PIRLS Literacy 2016 text across three languages', *Journal of Education* 87, 135–140. https://doi.org/10.17159/2520-9868/i87a07

Van Staden, S., 2006, *PIRLS of wisdom: The what, where, when and how of the International Reading Literacy Study in South Africa*, Centre for Evaluation and Assessment, Stellenbosch.

Von Davier, M., Mullis, I.V.S., Fishbein, B. & Foy, P., 2023, *Methods and procedures: PIRLS 2021 Technical Report*, Boston College TIMSS & PIRLS International Study Center, Boston.

Wry, E. & Mullis, I.V.S., 2023, 'Developing the PIRLS 2021 achievement instruments', in M. Von Davier, I.V.S. Mullis, B. Fishbein & P. Foy (eds.), *Methods and procedures: PIRLS 2021 technical report*, pp. 1.1–1.24, TIMSS & PIRLS International Study Centre, Boston.

Zumbo, B., 1999, *A handbook on the theory and methods of differential item functioning: Logistic regression modelling as a unitary framework for binary and Likert-type item scores*, Directorate of Human Resource Research and Evaluation, Ottawa, ON.