Journal of Education, 2022
Issue 87, http://journals.ukzn.ac.za/index.php/joe doi: http://dx.doi.org/10.17159/2520-9868/i87a07

Investigating the differential item functioning of a PIRLS Literacy 2016 text across three languages

Karen Roux

Department of Science, Mathematics and Technology Education, Faculty of Education, University of Pretoria, Pretoria, South Africa
karen.roux@up.ac.za
https://orcid.org/0000-0002-4181-1382

Surette van Staden

Department of Science, Mathematics and Technology Education, Faculty of Education, University of Pretoria, Pretoria, South Africa
surette.vanstaden@up.ac.za
https://orcid.org/0000-0002-5276-5705

Elizabeth J. Pretorius

Department of Linguistics and Modern Languages, College of Human Sciences, University of South Africa, Pretoria, South Africa
https://orcid.org/0000-0003-2137-1604

(Received: 16 November 2021; accepted: 7 June 2022)

Abstract

This study forms part of a larger study (Roux, 2020), which looked at the equivalence of a literary text across English, Afrikaans, and isiZulu from the Progress in International Reading Literacy Study (PIRLS). PIRLS is a large-scale reading comprehension assessment that assesses Grade 4 students' reading literacy achievement. PIRLS Literacy 2016 results for South African Grade 4 students indicated poor performance in reading comprehension, with approximately eight out of 10 Grade 4 students who could not read for meaning. Descriptive statistics led to the Rasch analysis, which was conducted using the South African PIRLS Literacy 2016 data. Even though the Rasch analysis indicated differential item functioning across the three languages for this specific passage, there was no universal discrimination against one particular language. By conducting differential item functioning, it was possible to determine whether the selected text had metric equivalence, in other words, whether the test questions were of similar difficulty across languages.

Keywords: differential item functioning, equivalence, large-scale studies, PIRLS Literacy 2016, reading comprehension, translation

Introduction and background

Since 2006, South Africa has participated in three cycles of the Progress in International Reading Literacy Study (PIRLS; cf. Howie et al., 2008, 2012, 2017). PIRLS is a large-scale international comparative study that assesses Grade 4 students' reading comprehension in 5-year cycles (Mullis & Martin, 2015). The PIRLS assessment comprised of narrative and informational texts. Students completed the PIRLS assessment in the language of learning and teaching (LoLT) in the Foundation Phase, for example, if the LoLT of the school was isiZulu, the Grade 4 students would have been tested in isiZulu. PIRLS is conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). The aim of PIRLS is to provide participating countries with reading comprehension trends across different cycles as well as the students' educational opportunities by collecting contextual information from the students' home environment, classroom practices, and school climate.

During each cycle, it was shown that South African students performed poorly in comparison with other participating countries. In the PIRLS 2016 study, the majority (78%) of South African Grade 4 students did not reach the Low International Benchmark (400 score points), which means that these students struggled with basic, literal questions and could not retrieve explicitly stated information in the text or make straightforward inferences about events or actions (Howie et al., 2017). Therefore, South African Grade 4 students are not moving beyond the literal understanding of the text and, as a result, not developing higher order reading comprehension skills (van Staden et al., 2019).

In the 2006 PIRLS study, South African Grade 4 students obtained a score approximately 250 score points below the international average of 500 (Howie et al., 2008). In the course of the next cycle, South Africa participated in pre-PIRLS 2011: an easier version of PIRLS 2011, which had a lower cognitive demand. Even though South Africa participated in pre-PIRLS 2011, the results were disappointing because the students obtained the lowest score (461 score points) amongst participating countries such as Botswana (463 score points) and Colombia (576 score points). In the next cycle of PIRLS, South African Grade 4 students participated in PIRLS Literacy 2016 (previously pre-PIRLS). Again, South Africa obtained the lowest score (320 score points) whereas other African countries such as Egypt and Morocco obtained 330 and 358 score points, respectively. This low achievement prompted national awareness of the importance of reading comprehension, with the President of South Africa including reading comprehension as a national priority (South African Government, 2019, 2020).[1]

1 It should be noted that a recent study by Gustafsson (2020) questioned the gains made between the 2011 and 2016 cycles by referring to the re-scaling of the data. Even though growth was underreported for the trends between 2011 and 2016, "it would be of importance to qualify where and what kind of growth we are seeing" (S. van Staden, personal communication, March 20, 2020). Based on the findings from Gustafsson's re-analysis, the trend data was recalculated by the IEA. Van Staden (2020, para. 19) stated that the gains made should be unpacked by conducting further analyses, specifically "into areas of the system that need improvement."

Based on the above, this article uses Rasch analysis, specifically differential item functioning (DIF), to examine the South African PIRLS Literacy 2016 results across English, Afrikaans, and isiZulu. The selection of these languages is based on performance, LoLT in South Africa, as well as the largest spoken language (isiZulu) in South Africa. This study only focuses on one of the passages examined as part of a larger study (Roux, 2020).

Literature review and conceptual framework

It is important to consider whether international large-scale assessments (ILSAs) such as PIRLS are equivalent across different languages and cultures, that is, whether the assessment items are understood similarly across all participants (Bundsgaard; 2019; Peña, 2007; Stubbe, 2011). Therefore, when assessments are designed and implemented across different countries, it is important that the same construct is measured and that achievement in the assessment only depends on the students' proficiency in the subject or topic area that is being measured.

The importance of equivalence

This article focuses on one of four considerations made by Peña (2007) when conducting cross-cultural and cross-lingual studies such as PIRLS. The four considerations highlighted by Peña (2007) include linguistic equivalence, metric equivalence, functional equivalence, and cultural equivalence.

Linguistic equivalence

Linguistic equivalence refers to the translation instructions for test instruments, which usually make use of the back-translation method (Chesterman, 2016). When a test instrument is translated and back-translated, the researchers should ensure that the source text (ST) is linguistically equivalent to that of the target text (TT). Consequently, the ST is translated and then back-translated by a second translator to ensure that words, sentences, and phrases are similar across both versions of the text (Behr, 2017). Thus, the aim of linguistic equivalence is to ensure that the linguistic meanings are the same across the ST and the TT (Peña, 2007).

Metric equivalence

The second consideration is metric equivalence, which concerns the difficulty of an item expressed in two different languages (Kim et al., 2003). The goal of metric equivalence is to make certain that the items used in a test are the same in terms of difficulty across different languages (Peña, 2007). When ILSAs such as PIRLS develop tests, these are usually in English and then translated into the different languages of the participating countries. It is important that when the test items are translated, that the source items' and target items' difficulty remain the same; the English items and, for example, the corresponding isiZulu items should have the same level of difficulty.

Functional equivalence

The third consideration entails functional equivalence. This provides evidence that the test instrument produces the same behaviour across different groups (Greenfield et al., 2006). Functional equivalence aims to ensure that there is a natural translation from the ST to the TT (Bermann & Porter, 2014). To put it another way, from a measurement perspective, the two versions of the test should behave similarly. The translator should keep in mind the receptors of the TT—meaning that the translated version of the text should be understood as if they were the receptors of the ST (Nida & Taber, 1969).

Cultural equivalence

The last consideration by Peña (2007) includes cultural equivalence. It considers how students interpret a test item that taps into the same cultural meaning for each cultural group (Chan & So, 2017). Basically, cultural equivalence aims to ensure that the meaning of the construct remains the same across different cultures and language groups (Peña, 2007). Each student who completes a test brings with them their own knowledge and understanding. Therefore, it may be difficult for ILSAs to achieve cultural equivalence given that each culture may perceive certain cultural aspects differently.

Equivalence of cross-cultural assessments refers to the similarity and comparability of an assessment across different language and cultural groups—and ensuring that students sampled from different populations, according to their LoLT, have equal opportunities to demonstrate their abilities (Peña, 2007). By focusing on metric equivalence and using DIF as evidence, it is possible to indicate whether there is item bias against a particular group of participants, in other words, whether there is measurement invariance.

Translation of international large-scale studies

Over the years, theories of translation and adaptation have been developed and adapted (Rios & Sireci, 2014; van de Vijver & Leung, 1997). In some part, the changes in educational assessment translation are due to the cross-cultural and cross-lingual nature of ILSAs (International Test Commission [ITC], 2017). Translation refers to linguistic discourse moving from a source language (SL) into the target language (TL). Furthermore, the act of translation requires the transfer of content and linguistic features from the SL to the TL. The translator should ensure that the translated version of the text is equivalent to that of the source text. Text characteristics such as the plot, setting, themes, characters, as well as the author's intent should be comparable and equivalent across the source and translated versions of text. According to Arffman (2013), when ILSAs translate their assessment instruments into multiple languages, these translated versions of the source texts should be "equivalent, or comparable, to each other" (p. 2).

Translation of PIRLS

PIRLS is a cross-cultural and cross-lingual study conducted in over 50 countries, worldwide. Each text that forms part of the study is translated into the different participating countries'

languages. The TIMSS & PIRLS International Study Center has, over the years, developed a rigorous translation verification procedure to ensure that the translations of the PIRLS texts are equivalent or comparable. Each participating country receives these guidelines to ensure that the texts are comparable across different languages and cultures (Martin et al., 2017). Ebbs and Wry (2017) stated:

> The ultimate goal of the translation and adaptation process was to create national versions of the PIRLS 2016 instruments that accommodate national languages and context while maintaining international comparability. (p. 7.1)

In terms of local translations, the aim of the South African PIRLS Literacy 2016 was to generate translations that were equivalent across the 11 official languages (Howie et al., 2017). After the South African National Research Coordinator received the international version of the texts, they were adapted to British English and then translated into the 10 remaining official languages. If any changes were made to the texts, the changes would have been recorded on the National Adaptation Forms in order to provide some form of quality assurance and to keep the changes minimal while, at the same time, acknowledging the different national contexts of each participating country.

Research question

One method to determine if there is item equivalence in the different versions is to perform item response theory analysis or Rasch analysis (cf. Andrich, 2011; Linacre, 2016). In their guidelines, the ITC (2017) acknowledged that it is possible that participants who write the translated and adapted versions may score lower or higher. Therefore, this article asks the following research question: "How did Grade 4 student performance differ across English, Afrikaans, and isiZulu languages on the Flowers on the Roof text?" The null hypothesis declares that the mean score of the English, Afrikaans, and isiZulu learners are equal ($_{\mu}$English = $_{\mu}$Afrikaans = $_{\mu}$isiZulu). However, if the null hypothesis is rejected, then an alternative hypothesis could be accepted (H_a = $_{\mu}$English ≠ $_{\mu}$Afrikaans ≠ $_{\mu}$isiZulu).

Research design and methods

Participants

The sample included 12,810 Grade 4 students from 293 schools who participated in the South African 2016 PIRLS study (Howie et al., 2017). In all schools, students were tested in the language of learning and teaching (LoLT) of the Foundation Phase (Grades 1 to 3); however, it needs to be acknowledged that the LoLT may not necessarily be the home language of the Grade 4 students. For the purposes of this study, a sub-sample (n = 761)[2] was selected of English (n = 323), Afrikaans (n = 186) and isiZulu (n = 252) students who completed the PIRLS Literacy 2016, Flowers on the Roof text.

[2] Fifty-nine Grade 4 learners were removed from the analysis due to extreme scores.

Data collection instruments

PIRLS Literacy 2016 included both achievement tests and background questionnaires. The achievement tests consisted of two types of texts, namely, reading for literary experience (narrative texts) and reading to acquire and use information (informational texts; Mullis & Martin, 2015). The background questionnaires gathered information regarding the students' educational environments such as the home, school, and classroom.

For the purposes of this article, attention is paid to the Flowers on the Roof text because it is one of the limited released texts used during PIRLS Literacy 2016. All other items from the PIRLS instruments are kept confidential.[3] The Flowers on the Roof text is a realistic fiction narrative text and revolves around two characters, Granny Gunn and a boy who is also the narrator of the story. The story focuses on intergenerational friendship and comfort, for example, making new friends and adding things around you so that you feel at home. Flowers on the Roof has 13 items that took the form of multiple-choice questions and constructed response questions. All multiple-choice items counted one point whereas the constructed response items ranged between one and three points. The majority of items (eight) entailed lower order skills: two items required students to find explicitly stated answers in the text, and six items needed students to make straightforward inferences based on the text. The remaining five items tested higher order skills such as interpreting information across the text and evaluating content (cf. Mullis et al., 2017, pp. 381–400).

The analysis of the PIRLS results is presented on a scale ranging from 0 to 1,000 with a fixed international centre point of 500, using plausible values derived from the three parameter item response theory model (Mullis et al., 2017). Therefore, an achievement score can be achieved, which is either above or below the international centre point.

Procedure

This study takes the form of a secondary analysis of the PIRLS Literacy 2016 data, specifically the Flowers on the Roof achievement data. The students' overall achievement was determined by using IDB-Analyzer, a software add-on for Statistical Package for the Social Sciences, created by the IEA because it can process large-scale data (Foy, 2018).

The descriptive statistics involved calculating the mean score for the Flowers on the Roof items for each of the three selected languages. Descriptive statistics were used to identify any differences in reading literacy achievement between English, Afrikaans, and isiZulu language sub-group responses.

After the initial exploration of the student results on Flowers on the Roof, differential item functioning (DIF) was conducted to address the research question. DIF is utilised to determine whether an item was much harder or easier for a group of respondents compared to a different group of respondents of equal ability, in other words, to determine whether

3 The PIRLS Literacy 2016 limited release passages and items are strictly confidential. Access to these passages and items in English and the translated versions must be granted by the IEA as well as the national research centre. The request form can be accessed here: https://www.iea.nl/publications/form/iea-permission-request-form

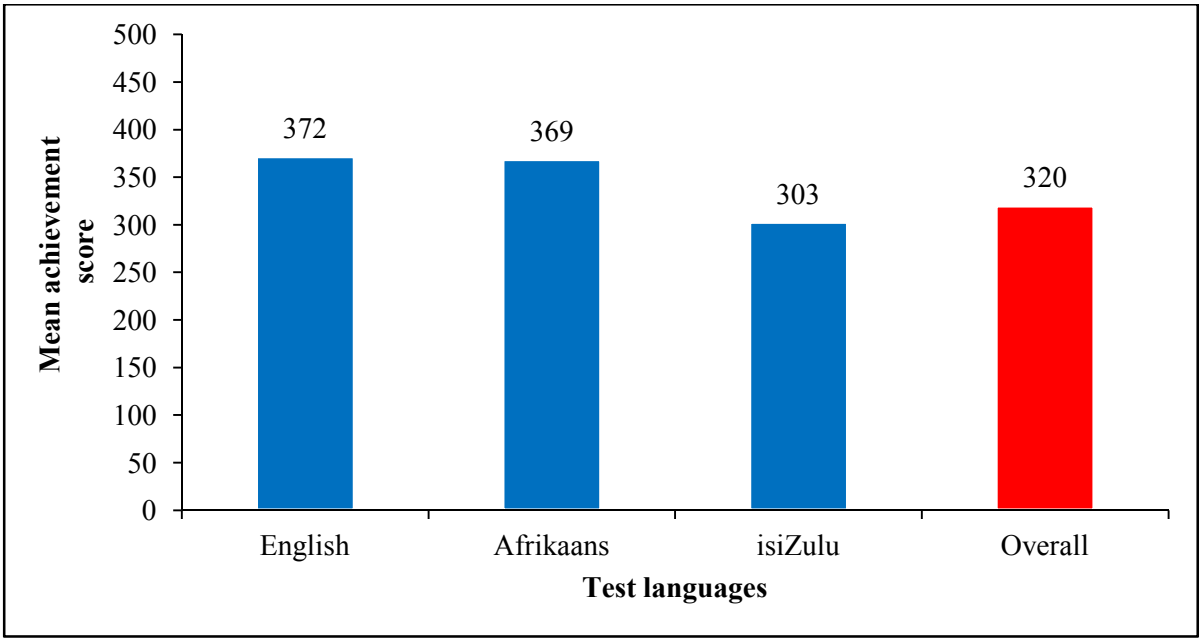
individual items functioned differently across different groups. DIF is a technique used to analyse, survey and, more importantly, test data that is conducted via Rasch measurement (Boone et al., 2014). Rasch analysis is a useful technique because it can detect differences in the item-level performance for different sub-groups of equal ability because person ability and item difficulty are included on the same scale. As part of this analysis, analysis of variance (ANOVA) statistics were produced to compare the means across the three languages. This study made use of the RUMM2030 software (Andrich et al., 2012) to analyse the PIRLS Literacy 2016 data.

Results

Descriptive findings

Figure 1 presents the overall South African Grade 4 student results of PIRLS Literacy 2016. This figure also indicates the English, Afrikaans, and isiZulu student results. The PIRLS scale ranges from 0 to 1,000 with an average of 500 (the centre point).

Figure 1
South African Grade 4 student achievement across selected languages (Source: Roux 2020, p. 123)



Overall, South African Grade 4 students obtained a mean score of 320 (*SE* = 4.4), which was the lowest among the participating countries (cf. Appendix 1 for percentiles and confidence interval levels across each language). Of the three selected languages for this study, the English students obtained the highest score (372 score points, *SE* = 14.4), followed by the Afrikaans (369 score points, *SE* = 13.4) and isiZulu (303 score points, *SE* = 4.4) students. No statistical difference was found between those students who wrote the PIRLS Literacy 2016 assessment in English and Afrikaans although both these languages achieved mean scores significantly higher than the isiZulu students ($p < 0.05$).

This article examines the Grade 4 English, Afrikaans, and isiZulu students' results for the Flowers on the Roof text. Table 1 presents the percentage of English, Afrikaans, and isiZulu students who correctly answered the 13 Flowers on the Roof items. Note that this table indicates the number of students who completed the test items and shows the percentage of students who were able to correctly answer each question. This analysis was conducted prior to the Rasch analysis to provide a snapshot of the specific items that were difficult for students.

Table 1
Percentage of Grade 4 students who correctly answered items in English, Afrikaans, and isiZulu (Source: Roux 2020, p. 129)

Item no.	English			Afrikaans			isiZulu		
	N Completed	*N* Correct	% Correct	*N* Completed	*N* Correct	% Correct	*N* Completed	*N* Correct	% Correct
1*	334	151	45**	196	59	30**	270	61	23**
2*	323	199	62	191	109	57	251	105	42**
3*	323	161	50	197	86	44**	260	92	35**
4*	333	176	53	198	82	41**	272	107	39**
5*	332	76	23**	198	85	43**	272	103	38**
6	326	65	20**	196	41	21**	252	11	4**
7	327	22	7**	197	19	10**	266	4	2**
8	320	49	15**	196	34	17**	265	10	4**
9	318	72	23**	195	20	10**	255	7	3**
10	313	99	32**	194	81	42**	243	71	29**
11*	321	108	34**	191	75	39**	248	52	21**
12	300	0	0**	181	0	0**	237	0	0**
13*	304	138	45**	182	64	35**	235	82	35**

*Indicates multiple-choice items. Remaining items are open-response questions.
**Less than 50% of the responses were correct.

Based on the results captured in Table 1, it is evident that the students found the text difficult. The text was especially difficult for the isiZulu students given that none of the items was correctly answered by at least 50% of the students. In terms of the Afrikaans students, 57% responded correctly to Item 2, while fewer than 50% responded correctly to the rest of the items. Similarly, 62%, 50%, and 53% of the English students respectively, answered Items 2, 3, and 4 correctly. Overall, Item 12 appeared to be the most difficult item because none of the students across the three languages was able to correctly answer it. To unpack the descriptive results, DIF was conducted to determine whether the items function in the same manner across the different languages.

Differential item functioning

The guiding question of this study asked how Grade 4 student performance differed across English, Afrikaans, and isiZulu languages on the Flowers on the Roof text. In order to gain a more comprehensive understanding of the aforementioned text, this article examined the individual item-fit as well the DIF results per item of the Flowers on the Roof text as a possible source of misft across the three languages (van Staden, 2018). The Bonferroni correction was selected for the ANOVA conducted with RUMM2030. This type of correction was used because some scholars have a concern regarding test of fit because it reduces the risk of Type I errors (Andrich & Marais, 2019).

The individual item-fit statistics of the text are in order of difficulty. Table 2 presents evidence of whether the items and the persons (Grade 4 students) link to the fit of the model because the assumption of the model declares that, as the person's ability increases, so should the chance of correctly responding to the more difficult items (Combrinck, 2019). However, when there is a lack of fit, it causes a violation of the assumption. In addition, Table 2 provides the chi-square value, which can reveal whether there is invariance across the trait (Pallant & Tennant, 2007).

Table 2
Individual item-fit statistics for Flowers on the Roof (Source: Roux 2020, p. 163)

Item	Difficulty	*SE*	Fit residual	Chi-Square	Probability
Flowers Item 2	–1,236	0,083	2,442	11,90	0,219
Flowers Item 4	–0,975	0,080	–3,456*	37,66	0,000**
Flowers Item 3	–0,837	0,082	–1,676	19,41	0,022
Flowers Item 13	–0,547	0,086	2,398	7,79	0,454
Flowers Item 1	–0,338	0,083	–2,742*	28,88	0,001**
Flowers Item 10	–0,260	0,087	1,780	33,17	0,000**
Flowers Item 5	–0,220	0,084	5,727*	52,72	0,000**
Flowers Item 11	–0,128	0,087	2,315	15,69	0,074
Flowers Item 9	0,260	0,063	–4,085*	33,41	0,000**
Flowers Item 7	0,909	0,076	–0,234	14,55	0,104
Flowers Item 12	0,964	0,079	–1,024	11,37	0,251
Flowers Item 6	0,989	0,107	–2,744*	18,05	0,035
Flowers Item 8	1,419	0,120	–3,530*	42,52	0,000**

*Fit residuals are indicated if above +2.5 or below –2.5.

**Bonferroni adjustment is 0.001282 for 13 items. All items smaller than the Bonferroni adjustment are indicated and are significant.

There is no clear pattern in terms of the PIRLS Process of Comprehension whether the students struggled more with higher order levels of comprehension such as interpreting information from across the text and evaluating content and textual elements (cf. Mullis & Martin, 2015). Five of the 13 items displayed a misfit that was significant. Item 5 was the only overfit item, meaning that the item discriminates a great deal between students. Items 4, 1, 9, and 8 displayed underfit, which means that the items do not adequately discriminate between the less able and more able students (van Staden, 2018). This finding indicates that these items have statistically significant chi-square probabilities, which shows they do not fit the model at the 5 per cent significance level.

Table 3 presents a summary of the Flowers on the Roof items that displayed DIF across the English, Afrikaans, and isiZulu languages. An ANOVA test is included in the DIF summary. This table offers evidence as to whether the mean scores of each language are indeed comparable.

Table 3
DIF summary for Flowers on the Roof text (Source: Roux 2020, p. 167)

Item	*F*-ratio	Probability
Flowers Item 2	1,278	0,279
Flowers Item 4	3,444	0,032
Flowers Item 3	0,406	0,666
Flowers Item 13	3,349	0,036
Flowers Item 1	8,644	0,000*
Flowers Item 10	6,661	0,001
Flowers Item 5	30,393	0,000*
Flowers Item 11	2,267	0,104
Flowers Item 9	19,500	0,000*
Flowers Item 7	2,597	0,075
Flowers Item 12	8,305	0,000*
Flowers Item 6	7,270	0,001*
Flowers Item 8	6,621	0,001
*Significant at the 5 per cent level (Bonferroni 0.001282)		

Table 3 demonstrates the ANOVA statistics. Results significant at the 5 per cent level are highlighted because the *p*-value is smaller than 0.05. Five of the Flowers on the Roof items

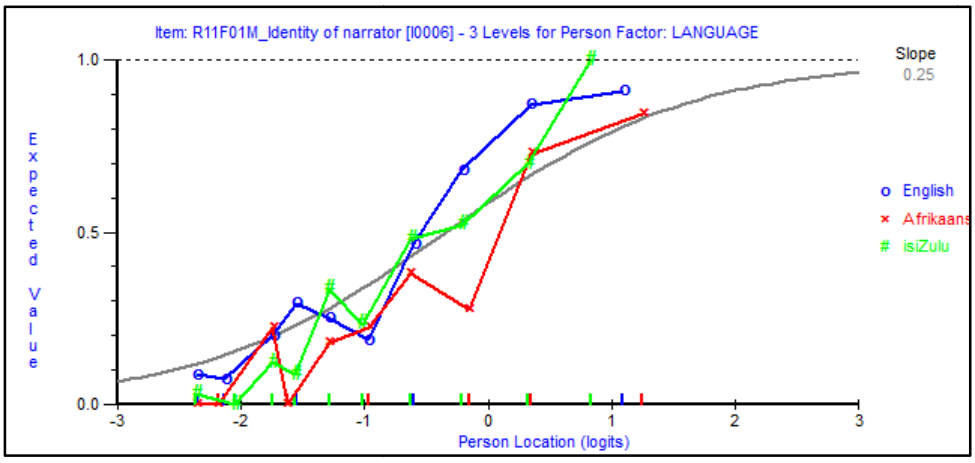
displayed differential functioning across the English, Afrikaans, and isiZulu sub-groups; these include Items 1, 5, 6, 9, and 12. Consequently, each of the aforementioned items was analysed further by looking at their item characteristic curves (ICC). The ICCs showed for each item, the differential functioning across languages as well as lower and upper class intervals (van Staden, 2018).

Flowers Item 1 was a multiple-choice question that required students to "Interpret and Integrate Ideas and Information" by asking:

> Who is telling the story?
>
> a. A granny.
> b. A child.* (correct answer)
> c. A doctor.
> d. A farmer.

Figure 2 indicates that there is extreme inconsistency across the lower class interval (between –3 and 0). Across all three languages, between the –3 and –2 person locations, the students had less than 10% probability of correctly responding to the item. Based on the figure, it would appear that item discriminated against students at the lower class interval. The distractor analysis provides evidence that students across the lower class interval were much more likely to select Distractor a (a granny) rather than the correct answer, namely, Distractor b (a child). A chi-square test of independence showed that there was a significant association between language and difficulty of the item, $\chi 2$ (9, $N = 761$) = 28.877, $p < .001$ which indicates that the distribution of distractors differs significantly from the expected model.

Figure 2
Flowers Item 1 characteristic curve (Source: Roux 2020, p. 169)



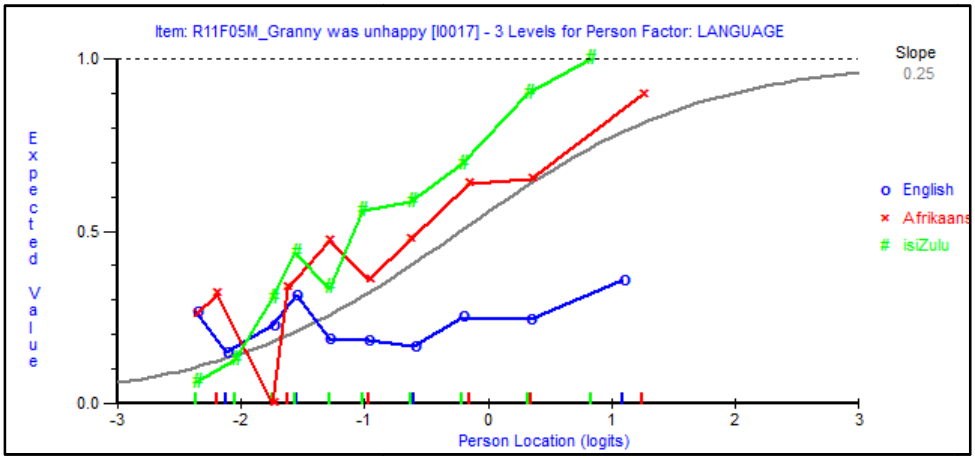
Flowers Item 5 is the next item that displayed DIF across the three languages that required students to make a straightforward inference. It was also a multiple-choice question asking the following:

Granny Gunn did not like the walls and windows in her new flat. Why else was she unhappy?

a. She was ill.
b. She missed her cat.
c. She did not like the balcony.
d. She felt homesick.* (correct answer)

Figure 3 presents the item characteristic curve for Flowers Item 5. The figure indicates extreme inconsistency across the three languages. Based on the item characteristic curve, the students who completed the test in English found the item considerably more difficult, and the English sub-group did not follow the expected model curve. The correct answer for Flowers Item 4 is Distractor d (she felt homesick), however, the students across the lower class interval were attracted by all three incorrect distractors. Moreover, the chi-square test of independence showed that there was a significant association between language and difficulty of the item, $\chi 2$ (9, N = 761) = 52.718, p < .001.

Figure 3
Flowers Item 5 characteristic curve (Source: Roux 2020, p. 171)

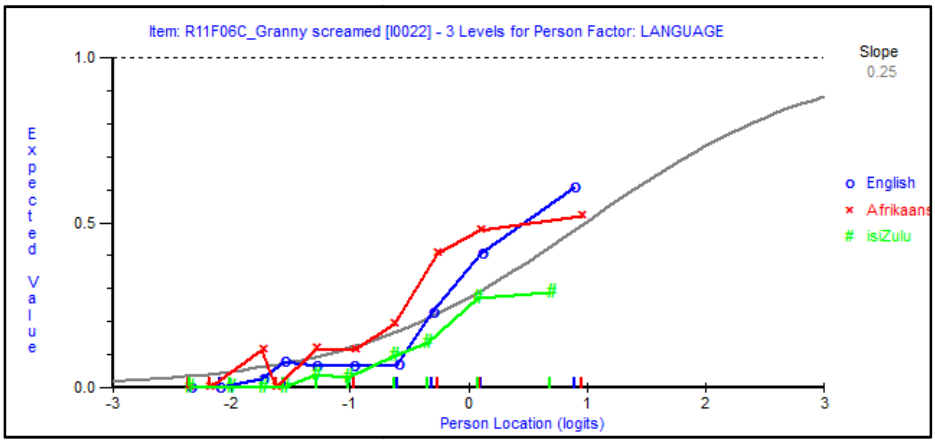


Flowers Item 6 was a constructed response type question, worth one mark, which required students to "Make Straightforward Inferences" from the text by asking:

Why did Granny Gunn scream when the cat jumped out of the window?

Figure 4 illustrates that the Grade 4 students found the item extremely difficult. The isiZulu students who completed this item between the –2.3 and –1.5 person locations had zero per cent chance of correctly responding to the item. Between the same person locations, the Afrikaans students experienced inconsistency whereas the English students had an approximate 10% chance of correctly responding to the item. Overall, the isiZulu students remained below the expected model curve across the lower and upper class intervals.

Figure 4
Flowers Item 6 characteristic curve (Source: Roux 2020, p. 176)

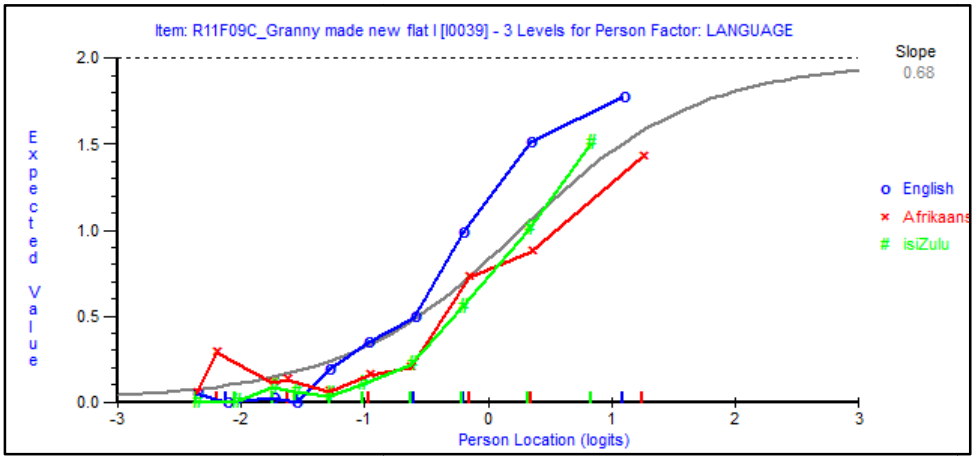


Flowers Item 9 took the form of a constructed type item for two marks. This question required students to "Focus on and Retrieve Explicitly Stated Information" from the text and asked the following:

> Write two ways in which Granny Gunn made her new flat feel like home.

Student responses had to include two actions of how Granny Gunn made her new flat feel like home. Figure 5 indicates inconsistency across the languages across both lower- and upper class intervals. Notably, the Afrikaans and isiZulu sub-groups were more often than not below the expected model curve.

Figure 5
Flowers Item 9 characteristic curve (Source: Roux 2020, p. 173)

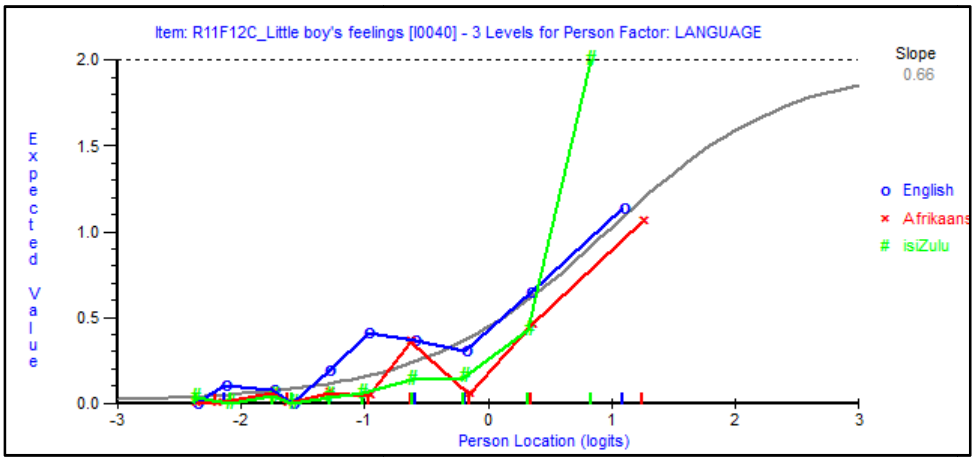


The last item, Flowers Item 12, was also a constructed response type item worth three marks. The item required students to "Interpret and Integrate Ideas and Information" from across the text by asking:

> What were the little boy's feelings about Granny Gunn when she first moved in and at the end of the story? Use what you have read to describe each feeling and explain why his feelings changed.

Even though the item carries three marks, none of the students was able to obtain full marks and therefore, the model discarded the three-mark parameter, using only the two-mark parameter. Figure 6 indicates that this item was exceptionally difficult for all students, especially for the students at the lower class interval (between –3 and 0). There is very little variation between –2.3 and –1.5 across the three languages. It appears that the students struggled to make the required inferences regarding the little boy's feelings about Granny Gunn and how his feelings changed over time. Peculiarly, the isiZulu sub-group at the 0.8 person location had a 100% chance of correctly responding to the item, in other words, obtaining at least two marks.

Figure 6
Flowers Item 12 characteristic curve (Source: Roux 2020, p. 174)



Discussion

The goal of this article was to gain insight into the South African Grade 4 students' reading comprehension performance, particularly on the Flowers on the Roof text that was one of the texts used during PIRLS Literacy 2016. In addition, this article aimed at providing evidence of item functioning for three South African languages given that the country has a multicultural and multilingual background.

The analysis by the current study illustrated the South African Grade 4 students' inability to correctly respond to both lower order and higher order reading skills. Of the five items that were detected as displaying some level of DIF, three items tested lower order skills such as finding explicitly stated information from the text or making straightforward inferences. The remaining two items tested the students' ability to interpret and integrate information across the text.

The item-fit statistics showed which items experienced over- or underfit across the three languages. Only five of the 13 items displayed misfit that was significant. This finding indicates that the items do not adequately discriminate between more and less able students (van Staden, 2018).

From the results of the item-fit statistics, DIF was conducted on the Flowers on the Roof items completed by the South African Grade 4 students. The ANOVA indicated that the five items functioned differently for the English, Afrikaans, and isiZulu students. These items included problematic responses between the three languages and provided possible evidence of measurement invariance across the Flowers on the Roof items. To gain a better understanding of these problematic items, item curves were conducted. In terms of the DIF and ICC, no clear pattern was observed given that there was no universal discrimination against any one language.

The translation of the text and items may have contributed to the poor performance because translation infelicities may occur between the source text and translated texts. However, as explained in the literature review, the IEA drafted a comprehensive guideline detailing aspects such as the selection criteria for translators, and the aim or goal of the translations. If any changes were made, these changes had to be meticulously added to the National Adaptation Forms to ensure that the changes did not affect the meaning or purpose of the text.

When conducting ILSAs such as PIRLS, it is important to ensure that the item difficulty remains the same, in other words, ensuring metric equivalence of the assessment (Peña, 2007). The PIRLS texts are developed in English, and then translated into the languages of the participating countries. In terms of the current study, Flowers on the Roof was provided to the participating countries in American English and then adapted into British English and translated into Afrikaans and isiZulu. It is important to keep in mind that the ST and TT may have different typological and orthographic features. English has an opaque orthography whereas African languages such as isiZulu have a transparent orthography (Spaull et al., 2020).

It is key that the item be at the same difficulty level across the different test languages and often, this gives rise to methodological complexities. According to Fischer et al. (2018), most research surrounding ILSAs focuses on pedagogic or systemic factors and not necessarily on translation issues. They contend that testing for measurement invariance is overlooked, although these differences may occur between countries due to the students' cultural background (Fischer et al., 2018). In his research, Stubbe (2011) investigated how different versions of a test instrument (PIRLS) function when translated into a single language. He found that even though only one language, German, was used, there was measurement invariance, especially with those items with differing translations. Furthermore, it was found that one version of the same test instrument was easier than the other two. This example of assessment of instrument translation highlights the importance of having metric equivalence across all assessment instruments.

Conclusion

The PIRLS Literacy 2016 results paint a bleak picture of the South African Grade 4 student results. This study analysed the South African Grade 4 results by focusing on the Flowers on the Roof text. By doing so, it was possible to determine whether there was measurement invariance across the English, Afrikaans, and isiZulu languages. One important characteristic of ILSAs such as PIRLS is to ensure that the test instruments are equal across different languages and cultures (cf. ITC, 2017; Peña, 2007). Another characteristic that should be taken into consideration includes the difficulty of the question items. Although this study only focused on metric equivalence, ILSAs should adhere to Peña's (2007) four considerations to assist in developing test instruments that are equivalent and that are similarly understood by the source and target readers.

The descriptive findings showed that, overall, the South African Grade 4 students struggled with most of the Flowers on the Roof items. Across the three languages, the majority of students were not able to provide correct answers for the items. A deeper analysis indicated that five of the 13 items displayed differential functioning. However, the findings of the Rasch analysis showed that the isiZulu students found Item 5 easier but Afrikaans and English students did not, and that English students found Item 9 less difficult than did Afrikaans or isiZulu students. There were also items that displayed extreme inconsistency across the three languages, which did not follow the model curve. If the Flowers on the Roof text had been unfair in one of the three languages, the above kind of variation would not have occurred.

Perhaps another opportunity may include examining the translations and metric equivalence of the remaining eight African languages. There may be translation infelicities for each of these languages that could partially explain these students' poor performance on the PIRLS 2016 test. One way of ensuring metric equivalence could be to develop a source language and target language glossary and word list (cf. Peña, 2007); however, this exercise may not be achievable because some languages may not have this kind of corpus linguistic data available. Yet, when assessments are developed for multiple languages and cultures, it could prove valuable to consider the different typologies of the languages given that word frequencies and word classes can be different (Ntshangase-Mtolo, 2009).

References

Andrich, D. (2011). *Rasch models for measurement*. SAGE Publications.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer.

Andrich, D., Sheridan, B. E., & Luo, G. (2012). *RUMM2030: Rasch unidimensional models for measurement*. RUMM Laboratory.

Arffman, I. (2013). Problems and issues in translating international educational achievement tests. *Educational Measurement: Issues and Practice, 32*(2), 2–14. https://psycnet.apa.org/doi/10.1111/emip.12007

Behr, D. (2017). Assessing the use of back translation: The shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology 20*(6), 573–584. https://doi-org.uplib.idm.oclc.org/10.1080/13645579.2016.1252188

Bermann, S., & Porter, C. (Eds.). (2014). *A companion to translation studies*. John Wiley & Sons.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Bundsgaard, J. (2019). DIF as a pedagogical tool: Analysis of item characteristics in ICILs to understand what students are struggling with. *Large-scale Assessments in Education, 7*(9), 1–14. https://doi.org/10.1186/s40536-019-0077-2

Chan, D. N. S., & So, W. K. W. (2017). Translation and validation of translation in cross-cultural research: Strategies used in a study of cervical cancer screening among ethnic minorities. *International Journal of Nursing Practice*, *23*(6), e12581. https://doi.org/10.1111/ijn.12581

Chesterman, A. (2016). *Memes of translation: The spread of ideas in translation theory* (revised ed.). John Benjamins.

Combrinck, C. (2019). *Rasch unidimensional measurement models: Analysing data in RUMM2030 hand-out*. Centre for Evaluation and Assessment.

Ebbs, D., & Wry, E. (2017). Translation and layout verification for PIRLS 2016. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures* (pp. 7.1–7.15). Boston College.

Fischer, J., Praetorius, A. K., & Klieme, E. (2018). The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*, *31*, 201–220. https://doi.org/10.1007/s11092-019-09295-7

Foy, P. (2018). *PIRLS 2016 user guide for the international database*. TIMSS & PIRLS International Study Center and the International Association for the Evaluation of Educational Achievement.

Greenfield, P. M., Trumbull, E., Keller, H., Rothstein-Fisch, C., Suzuki, L., & Quiroz, B. (2006). Cultural conceptions of learning and development. In P. A. Alexander, P. R. Pintrich, & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 675–694). Lawrence Erlbaum.

Gustafsson, M. (2020). *A revised PIRLS 2011 to 2016 trend for South Africa and the importance of analysing the underlying microdata* [Working papers 02/2020]. Stellenbosch University, Department of Economics. https://ideas.repec.org/p/sza/wpaper/wpapers337.html

Howie, S. J., Combrinck, C., Tshele, M., Roux, K., McLeod Palane, N., & Mokoena, G. M. (2017). *Progress in international reading literacy study 2016: South African children's reading literacy achievement*. University of Pretoria, Centre for Evaluation and Assessment.

Howie, S. J., van Staden, S., Tshele, M., Dowse, C., & Zimmerman, L. (2012). *PIRLS 2011: South African children's reading literacy achievement*. University of Pretoria, Centre for Evaluation and Assessment.

Howie, S. J., Venter, E., van Staden, S., Zimmerman, L., Long, C., du Toit, C., Scherman, V., & Archer, E. (2008). *PIRLS 2006 summary report: South African children's reading literacy achievement*. University of Pretoria, Centre for Evaluation and Assessment.

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.).

Kim, M., Han, H.-R., & Phillips, L. (2003). Metric equivalence assessment in cross-cultural research: Using an example of the Center for Epidemiological Studies-Depression Scale. *Journal of Nursing Measurement*, *11*, 5–18. https://psycnet.apa.org/doi/10.1891/jnum.11.1.5.52061

Linacre, J. M. (2016). *Winsteps Rasch measurement computer program user's guide*. https://www.winsteps.com/a/Winsteps-Manual.pdf

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017). *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework* (2nd ed.), TIMSS & PIRLS International Study Center.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. TIMSS & PIRLS International Study Center.

Nida, E. A., & Taber, C. R. (1969). *The theory and practice of translation* (4th ed.). Brill.

Ntshangase-Mtolo, P. (2009). *The translatability of English academic discourse into isiZulu with reference to the discourse of mathematics* [Master's thesis, University of KwaZulu-Natal, South Africa]. ResearchSpace. http://ukzn-dspace.ukzn.ac.za/handle/10413/1010

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, *46*, 1–18. https://doi.org/10.1348/014466506X96931

Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, *78*(4), 1255–1264. https://www.jstor.org/stable/4620701

Rios, J. A., & Sireci, S. G. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing*, *14*(4), 289–312. https://doi.org/10.1080/15305058.2014.924006

Roux, K. (2020). *Examining the equivalence of the PIRLS 2016 released texts in South Africa across three languages* [Doctoral thesis, University of Pretoria, South Africa]. UPSpace Institutional Repository. http://hdl.handle.net/2263/80509

South African Government. (2019). *President Cyril Ramaphosa: State of the Nation Address 2019.* https://www.gov.za/speeches/2SONA2019

South African Government. (2020). *President Cyril Ramaphosa: State of the Nation Address 2020.* https://www.gov.za/speeches/president-cyril-ramaphosa-2020-state-nation-address-13-feb-2020-0000

Spaull, N., Pretorius, E., & Mohohlwane, N. (2020). Investigating the comprehension iceberg: Developing empirical benchmarks for early-grade reading in agglutinating African languages. *South African Journal of Childhood Education*, *10*(1), article 773. https://doi.org/10.4102/sajce.v10i1.773

Stubbe, T. C. (2011). How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. *Educational Research and Evaluation*, *17*(6), 465–481. https://doi.org/10.1080/13803611.2011.630560

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. SAGE Publications.

van Staden, S. (2018). Exploring possible effects of differential item functioning on reading achievement across language subgroups: A South African perspective. In L. D. Hill & F. J. Levine (Eds.), *Global perspectives on education* (pp. 9.1–9.22). Routledge. https://doi.org/10.4324/9781351128421

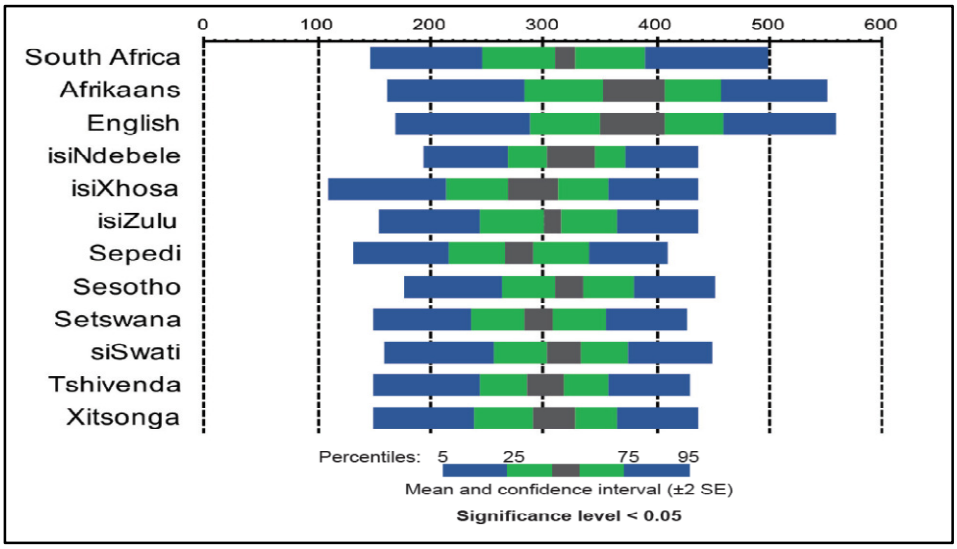
van Staden, S. (2020, April 6). International study shows where South Africa's education system needs more help. *The Conversation*. https://theconversation.com/international-study-shows-where-south-africas-education-system-needs-more-help-134448

van Staden, S., Combrinck, C., Roux, K., Tshele, M., & Palane, N. M. (2019). Moving beyond league table standings: How measures of opportunity to learn can inform educational quality and policy directives. *South African Journal of Childhood Education*, *9*(1), article 712. https://doi.org/10.4102/sajce.v9i1.712

Appendix 1

Figure 7 below depicts the distributions of South African Grade 4 students across the test languages, and includes the average scale score along with its 95 per cent confidence interval. The figure includes the 5th to 95th percentiles where the 25th to 75th percentiles comprise the middle of the students while the 5th and 95th show the extremes.

Figure 7
Comparison of South African PIRLS Literacy 2016 results per language (from Howie et al., 2017, p. 55).



This study only focused on three languages, namely, English, Afrikaans, and isiZulu. In terms of the variation of the mean scores, the greatest variation occurred in English and Afrikaans, which indicates that these two languages have a wider range in achievement compared to isiZulu, whereas isiZulu presented the least variation in terms of mean scores. Moreover, both English and Afrikaans at the 95th percentile achieved over 500 score points, whereas isiZulu obtained approximately 100 score points less (cf. Howie et al., 2017).